(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 August 2001 (09.08.2001)

PCT

(10) International Publication Number
WO 01/57065 A2

(51) International Patent Classification[7]: C07K 1/00

(21) International Application Number: PCT/GB01/00445

(22) International Filing Date: 2 February 2001 (02.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

0002492.7	3 February 2000 (03.02.2000)	GB
60/180,326	4 February 2000 (04.02.2000)	US
0016346.9	3 July 2000 (03.07.2000)	GB
0019362.3	7 August 2000 (07.08.2000)	GB

(71) Applicant *(for all designated States except US)*: **DIVERSYS LIMITED** [GB/GB]; 20 Park Crescent, London W1N 4AL (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RIECHMANN, Lutz** [DE/GB]; MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH (GB). **WINTER, Greg** [GB/GB]; MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH (GB).

(74) Agents: **MASCHIO, Antonio** et al.; D Young & Co, 21 New Fetter Lane, London EC4A 1DA (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/57065 A2

(54) Title: COMBINATORIAL PROTEIN DOMAINS

(57) Abstract: The present invention relates to a chimaeric folded protein domain comprising two or more sequence segments from parent amino acid sequences that are not homologous, wherein each of said sequence segments: is not designed or selected to consist solely of a single complete protein structural element and is not designed or selected to consist solely of an entire protein domain; and, in isolation, shows no significant folding at the melting temperature of the chimaeric protein; as well as to methods for the selection of such domains; and to a chimaeric folded protein domain comprising two or more sequence segments which share common sequences or sequences from common regions in the protein fold of their parent amino acid sequences, wherein the said common region of the fold is not designed or selected to consist solely of one or more complete protein structural elements, and wherein each of said sequence segments: is not designed or selected to consist solely of an entire protein domain; and wherein, in isolation, each said sequence segment shows no significant folding at the melting temperature of the chimaeric protein; as well as to methods for the selection of such domains.

Combinatorial Protein Domains

The present invention concerns the *de novo* synthesis of folded protein domains by the combinatorial rearrangement of sequence segments. The sequences of the segments may correspond directly to those of natural proteins, or be derived from those of natural proteins (for example by random or directed mutagenesis), or be derived by design based on the known structures of proteins. In particular, the invention makes use of combinatorial rearrangements of sequence segments which are not single entire structural elements of a natural protein and which, in isolation, show no significant folding.

The *de novo* design of proteins is typically based on structure predictions of predetermined amino acid sequences (Hecht 1994, Sauer 1996, Regan 1998). Partial randomisation is often introduced to allow for imperfection in the prediction algorithms. Resulting repertoires are screened or selected for stably folded structures. This approach has been successful for the design of helical structures like four helix bundles with stable and compact structures exhibiting free energies of unfolding of about 4 kcal/mol (Kamtekar *et al.*1993). More problematic has been the design of β-sheet proteins, where even the most recent attempts fell well short of natural β-sheet proteins regarding stability (Quinn *et al.*1994, Kortemme *et al.*1998, Alba *et al.*1999). Problems in the design of β-sheet structures are related to their dependence on backbone hydrogen bonds between different secondary structure elements, which are less well understood than the principles of helix formation (Hecht 1994). Repertoires of random protein sequences have also been screened for the occurrence of folded proteins. About 1% of members in a random library of Glu, Leu, Arg rich proteins exhibited some helix formation and cooperative unfolding but were unstable (Davidson & Sauer 1994).

Recently, new strategies to select stably folded proteins from repertoires of phage displayed proteins based on their resistance to proteolytic degradation have been used to improve the stability of natural proteins (Kristensen & Winter 1997, Sieber *et al.*1998, Finucane *et al.* 1999). Proteolytic degradation is usually restricted to unfolded proteins or highly flexible regions of folded proteins. Folded proteins are mostly resistant to proteases, because the proteolytic cleavage requires the polypeptide chain to adapt to the specific stereochemistry of the protease active site, and therefore to be flexible, accessible

and capable of local unfolding (Hubbard *et al.* 1994. Fontana *et al.* 1997). These methods have only been described for selection of proteins with point mutations; no element of combining sequences from different proteins is involved.

A theoretical approach to protein evolution via combinatorial rearrangement of defined. complete structural elements has been described (Bogarad & Deem 1999). The authors predict. using statistical algorithms. that rearrangement of a number of structural elements (such as helices. strands. loops. turns and others) will result in the generation of novel protein functions more rapidly than evolution by point mutation strategies alone. However. no allowance for the context dependence of structure is made. nor is any reference made to partial structural domains which possess no structural identity in isolation. Although some (rare) sequences will form structures in isolation. others can adopt a different structure in a different environment as evidenced by structural rearrangements following cleavage of some polypeptides by protease or on ligand binding. It is therefore not easy to define a structural element except in the context of the three dimensional structure of the protein in which it is embedded, and it is this definition that we have adopted here. Furthermore this paper does not show that it will be possible to undertake this process *in vitro*. or indicate exactly how to undertake such experiments.

Summary of the Invention

We have developed a different strategy for the creation and selection of novel protein domains which are capable of forming stably folded structures, and thus of identifying novel protein structural and functional elements.

The inventors have realised that as the structure of a "structural" element is dependent on context, single structural elements taken from one protein and appended to single structural elements taken from a second protein will not necessarily retain their original structure. Accordingly the inventors have not sought to restrict the segments to single complete structural elements. Furthermore the use of parts of structural elements can provide new structures that are not simply due to juxtaposition of existing structural elements. and the use of segments comprising multiple structural elements (and making packing interactions with each other) would be expected to be more stable than single

structural elements, and more likely to comprise a significant "nugget" of structure in the chimaeric domain.

Thus, the present invention exploits protein evolution by juxtaposition of sequence segments. "Sequence segments", as referred to herein, are amino acid sequences which are not designed or selected to consist solely of single and complete protein structural elements; and are not designed or selected to consist of a complete protein domain. The present invention is thus not directed to the juxtaposition of discrete and single elements of structure found in naturally-occurring or synthetic proteins, but with the juxtaposition of blocks of more than one structural element or with the creation of novel structural elements by the juxtaposition of sequences which, in isolation or in their parent environments, do not possess a discrete and complete structure.

Therefore, a "sequence segment" is an amino acid sequence which, in its parent environment, does not comprise a complete protein domain and is not encoded by one or more complete natural exons. Moreover, a "sequence segment", in its parent environment, does not form one or more discrete structural elements, but is either part of a structural element or, advantageously, is longer than a structural element. The sequence segment in isolation shows no significant folding at the melting temperature of the chimaeric protein; in other words, it possesses no independent structure in isolated form.

The "parent environment" of the sequence segment is the protein or polypeptide from which that segment is taken, in its folded state. This may be a natural protein, or an artificial polypeptide or protein. Preferably, the sequence segment is taken from an amino acid sequence which is longer than the sequence segment itself.

According to the present invention in a first configuration, the combinatorial rearrangement of protein sequence segments permits the selection of novel folded protein domains from combinatorial repertoires.

In a first aspect, therefore, the invention provides a chimaeric folded protein domain when derived from a repertoire of chimaeric proteins comprising two or more sequence segments derived from parent amino acid sequences that are not homologous.

Preferably, the parent amino acid sequences are derived from protein domains. The parent amino acid sequences may be natural, semi-synthetic or synthetic in origin. They may be derived by expression from genes or assembled by chemical synthesis.

Advantageously, the amino acid sequence segments are derived from proteins. In an advantageous embodiment, the proteins are selected from the group consisting of a naturally occurring protein, an engineered protein, a protein with a known binding activity, a protein with a known binding activity for an organic compound, a protein with a known binding activity for a peptide or polypeptide, a protein with a known binding activity for a carbohydrate, a protein with a known binding activity for a nucleic acid, a known binding activity for a hapten, a protein with a known binding activity for a steroid, a protein with a known binding activity for an inorganic compound, and a protein with an enzymatic activity.

As used herein, "amino acid" includes the 20 naturally-occurring amino acids, as well as non-naturally occurring amino acids and modified amino acids, such as tagged or labelled amino acids. As used herein, the term "protein" refers to a polymer in which the monomers are amino acids and are joined together through peptide or disulphide bonds. Preferably, "protein" refers to a full-length naturally-occurring amino acid chain or a fragment thereof, such as a selected region of the polypeptide that is of interest in a binding interaction, or a synthetic amino acid chain, or a combination thereof.

The sequence segments may be combined, in the chimaeric protein domain, by any appropriate means. Typically, the segments will be combined by recombinant DNA techniques and will thus be joined, in the recombinant protein, by peptide bonds. In alternative embodiments, the segments may be synthesised separately and subsequently joined. This may be achieved using covalent linkage, for instance peptide bonds, ester bonds or disulphide bonds, or non-covalent linkage. Advantageously, sequence segments according to the invention comprise one or more reaction groups for covalent or non-covalent linkage. For example, linkers capable of associating non-covalently, such as biotin/streptavidin, may be incorporated into the sequence segments to effect non-covalent linkage.

The repertoire from which the chimaeric protein domain is derived may be of substantially any size. Preferably, the repertoire comprises at least 10,000 individual protein domains; advantageously it comprises at least 1,000,000 protein domains; and most preferably, at least 100,000,000 protein domains.

The sequence segments may be any appropriate number of amino acids in length such that the combined length of the segments represents the length of a complete domain, which domains vary from as little as about 35 residues to several hundred residues in length.

In an advantageous aspect, the parent amino acid sequences are derived from the open reading frames of a genome or part thereof:

(a) wherein said reading frames are the natural reading frame of the genes; or

(b) wherein said reading frames are not the natural reading frame of the genes.

Sequences may thus be derived from ORFs present in a whole or substantially whole genome of an organism, or a part thereof, such as a group or family of genes, whether related by structure, function or evolution, or not related. The part of the genome may also consist of a single gene.

Sequences may moreover be derived from two or more genomes, from organisms of related or unrelated species.

The protein domains according to the invention are capable of folding due to the combination of two or more polypeptide segments which, in isolation, do not fold and do not define a single structural element in the parent protein.

Advantageously, the protein domains according to the invention are selected according to their resistance to proteolysis. This provides a useful means to isolate candidate domains from libraries; a selection procedure can be configured such that only proteolysis-resistant domains are selected from the libraries. Preferably, the proteolysis is carried out by exposure to a protease, such as thermolysin.

In a preferred embodiment, the protein domains according to the invention may be selected according to their activity. This may for example be a binding activity, for example in the case of immunoglobulin-type domains, or an enzymatic activity in the case of enzyme domains. Alternatively the protein domain may have the capacity to bind antibodies directed against the parent protein. Moreover, a screen for activity may be performed in addition to a selection on the basis of folding as determined by protease resistance. Such an approach is particularly advantageous where an initial selection on the basis of activity would be difficult or impossible to perform.

Moreover the invention concerns the juxtaposition of sequence fragments derived from non-homologous domains which share a similar polypeptide fold for at least part of the structure. We have observed that, in selections of protein domains according to the invention, sequence segments derived from parental protein domains having similar folds for at least part of their structures are juxtaposed in some of the novel chimaeric proteins. Accordingly, the present invention provides a chimaeric protein according to the first aspect of the invention, wherein the sequence segments originate from parent domains with similar polypeptide folds in at least part of the structure.

It has further been observed that, in selections of protein domains according to the invention, sequence segments derived from parental protein domains having entirely different folds for at least part of their structures are juxtaposed in other novel structures. Accordingly, the present invention provides a chimaeric protein domain comprising two or more sequence segments derived from parent amino acid sequences, wherein the sequence segments originate from parent domains with different polypeptide folds in at least part of the structure.

Moreover, in selections of protein domains according to the first configuration of the invention, sequence segments derived from the same protein domain may be observed to be juxtaposed to form novel structures. In some cases said sequence segments may comprise regions in common leading to a duplication of sequence in the chimaeric protein. However the common region does not consist of solely of one or more complete protein structural elements. Therefore it appears that duplication of amino acid segments or parts thereof, without regard to the presence of solely one or more complete structural

elements, can lead to the formation of stably folded structures. Such duplications comprise a second configuration of the invention.

As used herein, "regions in common" or "common regions" refers to regions which share sequence similarity or are of a similar fold. In this context, sequence similarity preferably refers to stretches of identical sequence of at least 10 amino acid residues: more preferably of at least 20 amino acid residues.

According to the second configuration of the invention, the combination of segments from homologous proteins, leading to equivalent regions from these homologous proteins being brought together in the same chimaeric protein, would also be expected to lead to the creation of stably folded structures. Regions, which are equivalent in homologous proteins, are identified by an alignment of their amino acid sequences. Indeed it is even possible to combine segments from non-homologous proteins which share a common fold (vide supra), to create stably folded chimaeric proteins from segments comprising a common region of the common fold in the parent proteins.

Said stably folded structures based on duplication of amino acid segments have been created as a product of the random shuffling of amino acid segments and were selected through proteolytic selection because of their stability. Duplication or indeed multimerisation performed in other non random ways have been previously reported, including for example by Hardies *et al.*1979 and Fire & Xu 1995. The inventors envisage that said methods for duplication and multimerisation may also be used for the duplication or multimerisation of amino acid segments to create novel and stably folded domains under the second configuration of the invention. Such stable domains may be selected and screened for in ways identical or similar to those in case of chimaeric domains derived from combinatorial shuffling.

Protein domains according to both configurations of the present invention may be created and selected by any suitable means. Preferred is combinatorial rearrangement of nucleic acid segments, for example in phage display libraries. Thus, the invention provides a chimaeric protein domain according to any foregoing aspect of the invention, fused to the

coat protein of a filamentous bacteriophage, said bacteriophage encapsidating a nucleic acid encoding the protein domain.

Moreover, both configurations of the invention provide a nucleic acid encoding a protein domain according to the invention as defined above.

In a further aspect of both configurations of the invention, the amino acid sequences of any chimaeric proteins may contain sequences designed to display epitopes for the vaccination against the parent protein of said amino acid sequences. For example, a chosen polypeptide segment from the coat protein of a virus, against which a vaccine is to be made, may be incorporated as a constitutive partner in a combinatorial library of amino acid sequences generated through the shuffling with one or more segments from another genetic source. Resulting chimaeric proteins will then comprise the segment of the viral coat protein in a variety of structural environments. By screening or selection, for example using antibodies from antisera raised against the virus, it is possible to identify those folded chimaeric proteins for which the viral sequence is displayed in a similar three dimensional configuration to the viral protein. Such stably folded proteins among these chimaeric constructs can be used for vaccination and elicit an immune response against the chimaeric protein which includes the viral amino acid segment. Vaccination with such a protein results in immunisation against the virus. One advantage compared to vaccination with the viral coat protein is that it is thereby possible to focus the immune response against one defined epitope of the virus, such as a neutralisation epitope.

It is also possible to vaccinate against defined epitopes of human proteins by the same strategy by combining a segment from a human protein with that from another source. The segment of non-human source should provide T-cell epitopes that will lead to an immune response against the human epitope. By way of example, it is possible to raise a blocking (IgG) antibody response against the portion of IgE that binds to the mast cell receptor. Such response is valuable, for example, in blocking asthma. This is achieved by construction of a chimaeric protein as follows. Firstly, segments from IgE are incorporated into chimaeric proteins by combination with a repertoire of non-human segments; secondly the proteins are screened or selected for binding to the mast cell receptor or to antibodies known to bind IgE at the critical site; thirdly the chimaeric

proteins with binding activities are used for immunisation. The IgE segments may be derived by random fragmentation of the IgE gene, or by using a segment already known to interact with the receptor. For immunisation it may be necessary to build in more potent T-cell epitopes into the non-human part, which can be achieved by making mutations in the non-human segment.

Preferably, therefore, the chimeric protein according to the invention comprises an epitope of a parent amino acid sequence. Advantageously, the epitope is a structural epitope.

Epitopes comprised in the chimeric proteins according to the invention, in a preferred embodiment, cross-react with antibodies raised against a parent amino acid sequence, or, advantageously, the folded parent protein.

In a further aspect of both configurations of the invention the segments may be derived entirely from human proteins. It is expected that these proteins will be less immunogenic in humans than foreign proteins as the sequences of the protein will be almost entirely human. Although such novel human proteins will be expected to differ in three dimensional structure from existing human proteins (and therefore to comprise novel B-cell epitopes), they will comprise T-cell epitopes derived from other human proteins (with the exception of the sequence flanking the join between segments). Such proteins, that are not immunogenic, or only weakly so, would be very suitable for therapeutic purposes or to avoid sensitisation in humans (for example enzymes in washing powders).

It is unlikely that in every respect that the chimaeric protein will mimic the three dimensional surface of the original protein in the region of target segment. This may be desirable in that it may allow the protein to adopt a conformation that has altered binding activities. For example, such proteins may be valuable as improved enzyme inhibitors.

Moreover the invention in either configuration provides for the creation of small domains that mimic part of the surface of a larger protein. One advantage of small domains is that it may more readily permit the three dimensional structure to be solved by X-ray

crystallography or NMR, and also at higher resolution. In turn this may facilitate the design of non-protein drugs based on the structure.

Further the invention in either configuration allows for the fusion of individual sequence segments juxtaposed in the chimaeric protein to additional, stably folded and complete protein domains. The function of additional domains may be to provide a means for selecting the chimaeric protein domains (see methods below). They may also serve to complement the chimaeric protein domain to perform a specific function. for example binding, immunogenicity or catalysis.

In the second configuration of the invention, the presence of at least two regions of the same sequence or similar (homologous) sequence in the chimaeric protein may permit the development of chimaeric proteins that bind to ligand at each of the two sites. This may be an advantage by giving improved "avidity" of binding where both heads can engage dimeric ligand (or other multimers), and also in providing two binding sites with different affinities, covering a larger dynamic range in binding to a ligand.

A further aspect of the first configuration of the invention relates to a method for selecting a protein domain according to the invention as defined above. Accordingly, the invention provides a method for preparing a protein domain according to the first aspect of the invention, comprising the steps of:

(a) providing a first library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences, said coding sequences not being selected or designed such as to solely encode a single and complete protein structural element or to encode a complete protein domain;

(b) providing a second library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences, said partner coding sequence not being selected or designed such as to solely encode a single and complete protein structural element or to encode a complete protein domain;

(c) combining the coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first and second libraries;

(d) transcribing and/or translating the contiguous coding sequences to produce the encoded protein domains;

(e) selecting the chimaeric protein domains which are able to adopt a folded structure or to fulfil a specific function.

Libraries according to the invention may be constructed such that sequences homologous to the partner coding sequence are excluded. For example, the libraries may be based on an artificial combination of solved structures, which means that the presence or absence of sequences homologous to the partner coding sequence can be controlled. However, if genomic libraries are used, it is possible that sequences homologous to the partner sequence may be present. In a preferred aspect, therefore, the method according to the invention further includes the steps of:

(f) analysing the sequence of the selected chimaeric protein domains to identify the origins of the sequence segments; and

(g) comparing the sequences of each of the parent amino acid sequences to identify whether the sequences of the parent amino acid sequences are non-homologous.

Similarly, it is possible to construct libraries comprising sequence segments derived from defined protein folds. However, if it is required to determine whether the isolated protein domain according to the invention is composed of sequence segments derived from parental domains having the same fold, the method according to the invention advantageously includes the step of:

(h) comparing the structures of each of the parent domains to identify whether they have same polypeptide folds in whole or in part.

In a further preferred aspect, the first configuration of the invention relates to the combination of a library of sequence segments with a unique partner coding sequence derived from a protein. The partner sequence is in this aspect provided as a unique sequence. Accordingly, steps (b) and (c) in the method according to the first configuration of the invention as set forth above may be modified such that:

(b) providing a partner coding sequence encoding a sequence segment derived from one protein, said partner coding sequence not being selected or designed such as to solely encode a single and complete protein structural element or to encode a complete protein domain:

(c) combining the library and partner coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first library and the partner coding sequence.

A further aspect of the second configuration of the invention relates to a method for selecting a protein domain, in which the individual sequence segments comprise common sequences. Accordingly, the invention provides a method for preparing a protein domain according to the first aspect of the invention, comprising the steps of:

(a) providing a first library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences, said coding sequences not being selected or designed to encode a complete protein domain;

(b) providing a second library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences, said partner coding sequence not being selected or designed such as to solely encode a single and complete protein structural element or to encode a complete protein domain;

(c) combining the coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first and second libraries;

(d) transcribing and/or translating the contiguous coding sequences to produce the encoded protein domains;

(e) selecting the chimaeric protein domains, which are able to adopt a folded structure or to fulfil a specific function;

and optionally:

(f) analysing the sequence of the selected chimaeric protein domains to identify the origins of the sequence segments; and

(g) comparing the sequences to identify whether they comprise common sequences.

Similarly, in a further aspect of the second configuration of the invention relates to a method for selecting a protein domain, in which the individual sequence segments comprise common regions from parent proteins with a common fold. However, if it is required to determine whether the isolated protein domain according to the invention is composed of sequence segments derived from parental domains having the same fold, the method according to the invention advantageously does not require step (g) above, but includes in its place the steps of:

(g) comparing the structures of the parent amino acid sequences to identify whether the parent proteins have a common fold; and

(h) identifying whether the segments comprise a common region of the common fold.

In a further preferred aspect, the second configuration of invention also relates to the combination of a library of sequence segments with a unique partner coding sequence derived from a protein. The partner sequence is in this aspect provided as a unique sequence. Accordingly, steps (b) and (c) in the method according to the second configuration of the invention as set forth above may be modified such that:

(b) providing a partner coding sequence encoding a sequence segment derived from one protein, said partner coding sequence not being selected or designed such as to solely encode a single and complete protein structural element or to encode a complete protein domain;

(c) combining the library and partner coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first library and the partner coding sequence.

Preferably, in the methods according to the both configurations of the invention the domains which are able to adopt a folded structure are selected by one or several methods selected from the group consisting of *in vivo* proteolysis, *in vitro* proteolysis, binding ability, functional activity and expression.

In a further aspect, an amino acid sequence of any chimaeric proteins produced through combinatorial shuffling according to both configurations of the invention may be mutated or altered after the original juxtaposition of the parent amino acid sequences. Such changes may be introduced by any of the following methods:

(a) designing and introducing specific or random mutations at predefined positions within the gene of the chimaeric protein;

(b) deleting nucleotides within the gene of the chimaeric protein so as to delete amino acid residues;

(c) inserting nucleotides within the gene of the chimaeric protein so as to insert amino acid residues

(d) appending nucleotides to the gene of the chimaeric protein so as to append amino acid residues;

(e) randomly introducing mutations in all or part of the gene encoding the chimaeric protein through recombinant DNA technology;

(f) randomly introducing mutations in the gene of the chimaeric protein through propagation in mutating cells;

(g) introducing derivatives of natural amino acid during chemical synthesis;

(h) chemically derivatising amino acid groups after synthesis;

(i) multimerising the chimaeric proteins through concatenation of two or more copies of the gene in a single open reading frame;

(j) multimerising the chimaeric proteins through covalent linkage of two or more copies of the chimaeric protein domain after translation;

(k) multimerising the chimaeric proteins through fusion to a multimeric partner.

Any said changes may improve the stability or the function of the chimaeric protein. For example, said changes may be aimed to meet predicted structural requirements within the combined segments advantageous for the formation of specific polypeptide folds or to introduce specific amino acid sequences to fulfil a desired function. An example of such improvements is given in Example 14 in the Experimental section.

The invention moreover encompasses further optimisation of the regions of N- and C-termini of recombined amino acid segments. Both their joining and end regions as part of

a chimaeric protein are conceivably not optimised as far as stability and/or function of the chimaeric protein are concerned. Natural proteins, which may have been created through a recombinatorial event, are subsequently optimised through (point) mutational events and Darwinian selection. This process may be mimicked *in vitro* for chimaeric protein as defined herein, for example using the above listed methods (including mutation, deletion and/or addition of amino acid residues).

Chimaeric proteins containing such improvements may be identified by one or more methods used for the selection and screening of the original combinatorial library. It may further be advantageous to produce any selected chimaeric protein domains in a multimerised form, for example to increase stability through interdomain interaction or improve binding to a ligand through avidity effects.

Brief Description of the Figures

Figure 1. Proteolysis of selected phages and chimaeric proteins. (a) ELISA for barstar binding of phages 1c2 (squares), 1b11 (circles), 1g6 (diamonds) and csp/2 (triangles) before and after trypsin/thermolysin treatment at different temperatures. (b) SDS-PAGE of proteins His-1c2, His-1b11 and His-1g6 before and after treatment with trypsin, thermolysin and chymotrypsin at 25°C.

Figure 2. Circular dichroism and thermodenaturation of chimaeric proteins. (a) Circular dichroism spectra of His-1c2 (upper trace) and His-2f3 (lower trace) at 20°C. (b) Ellipticity of His-1c2 (at 205 nm; upper trace) and His-2f3 (at 223 nm; lower trace) at different temperatures.

Figure 3. Nuclear magnetic resonance analysis of chimaeric proteins. 1D-^{1}H-NMR spectra of His-2f3 recorded (a) at 25°C in H_2O and (b) after incubation for 24 hours at 25°C in D_2O. 1D-^{1}H-NMR spectra of His-1c2 recorded at 30°C (c) in H_2O and (d) after incubation for 24 hours at 25°C in D_2O. 2D-^{1}H-NOESY spectrum of His-1c2 recorded at 30°C (e) in H_2O.

Figure 4. Biotin-CspA ELISA. A rabbit anti-CspA antiserum was incubated with varied amounts of soluble His-CspA, His-1c2, His-2f3, His-1b11 or lysozyme (as a negative control) before binding to biotinylated CspA immobilised on Streptavidin coated ELISA well. Bound rabbit antisera were detected with a HRP-conjugated goat anti-rabbit IgG antiserum.

Detailed Description of the Invention

The present invention relates to chimaeric, folded protein domains. In the context of the present invention, "folded" means that the protein domains concerned are capable of adopting, or have adopted, a stable tertiary structure. Stability in this context may be defined as the conformational stability of the protein, which is the difference in free energy between the folded and unfolded conformations under physiological conditions; the higher this value, the greater the energy required to unfold the protein, and thus the greater the stability of the folded structure. A quantitative measure of this conformational stability of proteins, the Gibbs free energy of folding, can be determined from reversible thermodynamics. Proteins undergo order-disorder transitions, which are detectable in differential scanning calorimetry (DSC) profiles of specific heat vs. temperature.

Preferably, the free energy of folding possessed by a protein domain according to the invention is 1.6 kcal/mol or higher; advantageously, it is 3 kcal/mol or higher; and most preferably it is 5 kcal/mol or higher.

Folded proteins which form stable structures are known to be resistant to proteolysis. Thus, the invention provides for the selection of folded protein domains in accordance with the present invention using protease enzymes, which cleave and preferably eliminate unstable or unfolded domains. "Folded" may therefore be defined in terms of resistance to proteolysis under assay conditions. Exemplary conditions are set forth in the examples below.

Sequence segments according to the invention are segments of natural protein sequence, which occurs in naturally-occurring proteins, or artificial segments of sequence modelled on the sequence or structure of naturally-occurring proteins. The sequence segments may

be between 10 and 100 amino acids in length, or longer; preferably between 15 and 50 amino acids in length; and advantageously between 20 and 45 amino acids in length; or, where nucleic acids are concerned, the necessary length to encode such amino acid sequences.

Sequence segments according to the invention are derived from parental protein domains which are not homologous.

The term "parent amino acid sequences" (or "parental amino acid sequences") refers to any amino acid sequences encoded by open reading frames within DNA sequences, which form the source of the cloned DNA segments as part of the combinatorial libraries as outlined in the claims. Said reading frames may be part of the original reading frame of genes, of shifted reading frames or of the reverse strand of genes. They may also form part of intragenic regions, which are not known to encode a protein. Originating genes may be natural or synthetic.

As outlined in the introduction, the term homology between two or more proteins or proteins domains can refer to a similarity or identity of both their amino acid sequences and their structural fold. For the present purposes, the term homology shall solely refer to the degree of identity between two parent amino acid sequences.

Homologous amino acid sequences have 35% or greater identity (e.g., at least 40% identity, 50% identity, 60% identity, 70% identity, or at least 80% identity, such as at least 90% identity, or even at least 95% identity, for instance at least 97% identity). Homologous nucleic acid sequences are nucleic acid sequences which encode homologous polypeptides, as defined. Actual nucleic acid sequence homology/identity values can be determined using the "Align" program of Myers & Miller 1988, ("Optimal Alignments in Linear Space") and available at NCBI. Alternatively or additionally, the term "homology", for instance, with respect to a nucleotide or amino acid sequence, can indicate a quantitative measure of homology between two sequences. The percent sequence homology can be calculated as $(N_{ref} - N_{dif})*100/N_{ref}$, wherein N_{dif} is the total number of non-identical residues in the two sequences when aligned and wherein N_{ref} is the number of residues in one of the sequences. Hence, the DNA sequence AGTCAGTC

will have a sequence similarity of 75% with the sequence AATCAATC (N_{ref} = 8; N_{dif} = 2). Alternatively or additionally, "homology" with respect to sequences can refer to the number of positions with identical nucleotides or amino acids divided by the number of nucleotides or amino acids in the shorter of the two sequences wherein alignment of the two sequences can be determined in accordance with the Wilbur and Lipman algorithm (Wilbur & Lipman 1983), for instance, using a window size of 20 nucleotides, a word length of 4 nucleotides, and a gap penalty of 4, and computer-assisted analysis and interpretation of the sequence data including alignment can be conveniently performed using commercially available programs (e.g., Intelligenetics™ Suite, Intelligenetics Inc. CA). When RNA sequences are said to be similar, or have a degree of sequence identity or homology with DNA sequences, thymidine (T) in the DNA sequence is considered equal to Uracil (U) in the RNA sequence.

RNA sequences within the scope of the invention can be derived from DNA sequences, by thymidine (T) in the DNA sequence being considered equal to Uracil (U) in RNA sequences.

Additionally or alternatively, amino acid sequence similarity or identity or homology can be determined using the BlastP program (Altschul *et al.* 1997) and available at NCBI. The following references (each incorporated herein by reference) provide algorithms for comparing the relative identity or homology of amino acid residues of two proteins, and additionally or alternatively with respect to the foregoing, the teachings in these references can be used for determining percent homology or identity: Needleman & Wunsch (1970); Smith & Waterman (1981); Smith *et al.*(1983); Feng & Dolittle (1987); Higgins & Sharp (1989); Thompson *et al.*(1994); and Devereux *et al.*(1984).

The invention contemplates the recombination of sequence segments which are derived from parental proteins with similar folds. In this context, "similar" is not equivalent to "homologous". Indeed, similar folds have been shown to arise independently during evolution. Such folds are similar but not homologous.

A "protein structural element" is an amino acid sequence which may be recognised as a structural element of a protein domain. Preferably, the structural element is selected from

the group consisting of an α-helix, a β-strand, a β-barrel, a parallel or antiparallel β-sheet, other helical structures (such as the 3_{10} helix and the pi helix), and sequences representing tight turns or loops. Advantageously, the structural element is an α-helix or a β-strand, sheet or barrel.

In a preferred embodiment, the folded protein domains according to the present invention are constructed from sequence segments which do not comprise only a single structural element; rather, they comprise less than a single structural element, or more than a single structural elements or parts thereof.

In accordance with the present invention, the sequence segments used are not designed or selected to comprise only such single elements; in other words, they may comprise more than a single structural element, or less than a single structural element. This may be achieved through the use of substantially random sequence segments in constructing a library according to the invention. For example, sonicated genomic or cDNA or segments produced by random PCR of DNA may be used. Advantageously, the DNA fragments are between 100 and 500 nucleotides in length.

The sequence segments used in accordance with the present invention are unable to fold significantly in isolation; that is, they do not contain sufficient structural information to form a folded protein domain unless they are combined with another sequence segment in accordance with the present invention. The inability to fold significantly may be measured by susceptibility to protease digestion, for example under the conditions given in the examples below, or by measurement of the free energy of folding .

Proteolysis may be carried out using protease enzymes. Suitable proteases include trypsin (cleaves at Lys, Arg), chymotrypsin (Phe, Trp, Tyr, Leu), thermolysin (small aliphatic residues), subtilisin (small aliphatic residues), Glu-C (Glu), Factor Xa (Ile/Leu-Glu-Gly-Arg), Arg-C (Arg) and thrombin. Advantageously, since the combination of random polypeptide sequence segments cannot be guaranteed to generate a precise cleavage site for a particular protease, a broad-spectrum protease capable of cleaving at a variety of sites is used. Trypsin, chymotrypsin and thermolysin are broad-spectrum proteases useful in the present invention.

The ability of a protein domain to fold is also associated with its function. Accordingly, the invention provides for the selection of folded protein domains by functional assays.

In the case of immunoglobulins or other polypeptides capable of binding, such assays may be performed for binding activity according to established protocols: however, where binding is only transitory, the selection may be performed on the basis of function alone. Suitable methodology is set forth, for example, in International patent applications PCT/GB00/00030 and PCT/GB98/01889. Such techniques are useful for the selection of novel or improved enzymes produced by combinatorial rearrangement according to the present invention.

The invention also provides for screening for activity after selection according to protease resistance. This allows protein domains which have been selected according to their ability to fold to be screened for any desired activity. Since the repertoire sizes are more limited, as a result of the selection by proteolysis, the screening step can be conducted more easily (for example, in a multiwell plate).

The libraries of the present invention may be created by any suitable means in any form. As used herein, the term "library" refers to a mixture of heterogeneous polypeptides or nucleic acids. The library is composed of members, each of which has a unique polypeptide or nucleic acid sequence. To this extent, *library* is synonymous with *repertoire*. Sequence differences between library members are responsible for the diversity present in the library. The library may take the form of a simple mixture of polypeptides or nucleic acids, or may be in the form organisms or cells, for example bacteria, viruses, animal or plant cells and the like, transformed with a library of nucleic acids. Typically, each individual organism or cell contains only one member of the library. In certain applications, each individual organism or cell may contain two or more members of the library. Advantageously, the nucleic acids are incorporated into expression vectors, in order to allow expression of the polypeptides encoded by the nucleic acids. In a preferred aspect, therefore, a library may take the form of a population of host organisms, each organism containing one or more copies of an expression vector containing a single member of the library in nucleic acid form which can be expressed to

produce its corresponding polypeptide member. Thus, the population of host organisms has the potential to encode a large repertoire of genetically diverse polypeptide variants.

A number of vector systems useful for library production and selection are known in the art. For example, bacteriophage lambda expression systems may be screened directly as bacteriophage plaques or as colonies of lysogens, both as previously described (Huse *et al.*(1989); Caton & Koprowski (1990); Mullinax *et al.*(1990); Persson *et al.*(1991) and are of use in the invention. Whilst such expression systems can be used to screening up to 10^6 different members of a library, they are not really suited to screening of larger numbers (greater than 10^6 members). Other screening systems rely, for example, on direct chemical synthesis of library members. One early method involves the synthesis of peptides on a set of pins or rods, such as described in WO84/03564. A similar method involving peptide synthesis on beads, which forms a peptide library in which each bead is an individual library member, is described in U.S. Patent No. 4,631,211 and a related method is described in WO92/00091. A significant improvement of the bead-based methods involves tagging each bead with a unique identifier tag, such as an oligonucleotide, so as to facilitate identification of the amino acid sequence of each library member. These improved bead-based methods are described in WO93/06121.

Another chemical synthesis method involves the synthesis of arrays of peptides (or peptidomimetics) on a surface in a manner that places each distinct library member (e.g., unique peptide sequence) at a discrete, predefined location in the array, or the spotting of pre-formed polypeptides on such an array. The identity of each library member is determined by its spatial location in the array. The locations in the array where binding interactions between a predetermined molecule (e.g., a receptor) and reactive library members occur is determined, thereby identifying the sequences of the reactive library members on the basis of spatial location. These methods are described in U.S. Patent No. 5,143,854; WO90/15070 and WO92/10092; Fodor *et al.*(1991); and Dower & Fodor (1991).

Of particular use in the construction of libraries of the invention are selection display systems, which enable a nucleic acid to be linked to the polypeptide it expresses. As used

herein, a selection display system is a system that permits the selection, by suitable display means, of the individual members of the library.

Any selection display system may be used in conjunction with a library according to the invention. Selection protocols for isolating desired members of large libraries are known in the art, as typified by phage display techniques. Such systems, in which diverse peptide sequences are displayed on the surface of filamentous bacteriophage (Scott & Smith (1990), have proven useful for creating libraries of antibody fragments (and the nucleotide sequences that encoding them) for the *in vitro* selection and amplification of specific antibody fragments that bind a target antigen. The nucleotide sequences encoding the V_H and V_L regions are linked to gene fragments which encode leader signals that direct them to the periplasmic space of *E. coli* and as a result the resultant antibody fragments are displayed on the surface of the bacteriophage, typically as fusions to bacteriophage coat proteins (e.g., pIII or pVIII). Alternatively, antibody fragments are displayed externally on lambda phage capsids (phagebodies). An advantage of phage-based display systems is that, because they are biological systems, selected library members can be amplified simply by growing the phage containing the selected library member in bacterial cells. Furthermore, since the nucleotide sequence that encode the polypeptide library member is contained on a phage or phagemid vector, sequencing, expression and subsequent genetic manipulation is relatively straightforward.

Methods for the construction of bacteriophage antibody display libraries and lambda phage expression libraries are well known in the art (McCafferty *et al.*(1990); Kang *et al.*(1991); Clackson *et al.*(1991); Lowman *et al.*(1991); Burton *et al.*(1991); Hoogenboom *et al.*(1991); Chang *et al.*(1991); Breitling *et al.*(1991); Marks *et al.*(1991); Barbas *et al.*(1992); Hawkins & Winter (1992); Marks *et al.*(1992); Lerner *et al.*(1992), incorporated herein by reference).

Other systems for generating libraries of polypeptides or nucleotides involve the use of cell-free enzymatic machinery for the *in vitro* synthesis of the library members. For example, *in vitro* translation can be used to synthesise polypeptides as a method for generating large libraries. These methods which generally comprise stabilised polysome complexes, are described further in WO88/08453, WO90/05785, WO90/07003,

WO91/02076. WO91/05058. and WO92/02536. Alternative display systems which are not phage-based. such as those disclosed in WO95/22625 and WO95/11922 (Affymax) use the polysomes to display polypeptides for selection. These and all the foregoing documents are incorporated herein by reference.

In order to produce libraries of sequence segments in accordance with the present invention. PCR amplification is advantageously employed. Where a defined partner sequence is used, one PCR primer may be deigned to anneal specifically with the partner sequence; for random libraries, general random PCR primers may be used. The resulting fragments are joined by restriction and ligation and cloned into suitable vectors. Although the ligation of two sequence segments is described below. the invention encompasses the ligation of three or more sequence segments. any of which may be the same or different. such as to mirror a multiple cross-over event.

The invention is further described. for the purpose of illustration. in the following experimental section.

Example 1

Preparation of a repertoire of chimaeric proteins comprising two sequence segments

A repertoire of genes encoding chimaeric proteins, which comprise the N-terminal 36 residues of the *E. coli* cold shock protein (CspA) and a C-terminal polypeptide sequence encoded by randomly created fragments of the *E. coli* genome, was prepared. CspA comprises 70 residues and forms a stable β-barrel (Schindelin *et al.*1994). Its N-terminal 36 residues comprise the first three strands of its six stranded β-barrel and are unable to fold when expressed alone as they are degraded in the *E. coli* cytoplasm.

The gene fragment encoding the first 36 residues of CspA was complemented with fragmented DNA from The *E. coli* genome around 140 base pairs in size. DNA fragments were created by random PCR amplification using genomic *E. coli* DNA as a template. Resulting chimaeric genes were inserted between the coding regions for the infection protein p3 and an N-terminal tag. a stable but catalytically inactive mutant of the RNase

barnase. as a single continuous gene on a phagemid vector for protein display on filamentous phage.

In the resulting genomic library ($1.0x10^8$ members) an opal (TGA) stop codon was incorporated at the 3' end of the chimaeric gene in 60% of clones with the remainder containing the Gly-encoding GGA codon in this position. The partial incorporation of the TGA codon at the 3' end of the chimaeric genes was achieved through the use of two different PCR primers (XTND and NOARG) in the PCR amplifications of the *E. coli* gene fragments. The transfer-RNA^{Trp} can decode TGA with an efficiency of up to 3% (Eggertsson & Söll 1988) leading to sufficient display of the barnase-chimera-p3 fusion on the phage but avoiding folding related. toxic effects. Phages displaying this repertoire were prepared using the helper phage KM13, which contains a modified fd gene 3 encoding a trypsin-sensitive p3 due to a modified sequence (Kristensen & Winter 1997). to reduce infectivity due to helper phage encoded p3 molecules.

Example 2

Preparation of a repertoire of chimaeric proteins comprising two sequence segments with common sequences

In a second "plasmid-derived" library the N-terminal CspA gene fragment was complemented with DNA fragments of around 140 base pairs created by random PCR amplification using as the PCR template a 3.6 kb plasmid containing the wild type CspA gene. Resulting chimaeric genes were again inserted as a fusion between the coding regions for the infection protein p3 and an N-terminal tag, a stable but catalytically inactive mutant of the RNase barnase, on a phagemid vector for protein display on filamentous phage.

In the plasmid-derived library ($1.7x10^8$ members) an opal (TGA) stop codon was constitutively introduced at the 3' end of the chimaeric gene in all clones. Phages displaying this repertoire were prepared using the helper phage KM13, which contains a modified fd gene 3 encoding a trypsin-sensitive p3 due to a modified sequence

(Kristensen & Winter 1997), to reduce infectivity due to helper phage encoded p3 molecules.

Example 3

Proteolytic selection of combinatorial libraries

To select stably folded chimaeric proteins from the repertoires of barnase-chimaera-p3 fusions described in Examples 1 and 2, the phage-displayed libraries were select for proteolytic stability in three rounds through treatment at 10°C with the proteases trypsin (specific for peptide bonds containing Arg or Lys in the P_1 position) and thermolysin (specific for bonds containing a amino acid with an aliphatic side chain in the $P_{1'}$ position) followed by capture on barstar, elution, infection and regrowth.

After the first round of selection, $2x10^4$ and $6x10^2$ of 10^{10} proteolytically treated phages were eluted from a single barstar coated microtitre plate well in case of the plasmid-derived library and the genomic library, respectively. When protease treatment is omitted $5x10^6$ phages can be eluted indicating that the vast majority of unselected phages did not display a stably folded chimera protein fused between barnase and p3. The number of phages rescued after two and three rounds of selection increased to $2x10^5$ for the plasmid-derived library and to $2x10^3$ and $4x10^4$ for the genomic library.

Selected phages were grown up individually, bound to immobilised barstar, treated *in situ* with trypsin and thermolysin at 10°C and resistance was measured through detection of bound (and therefore resistant) phage in ELISA. For the plasmid-derived library 27 of 64 phages (42%) retained 80% or more of their barstar binding activity after protease treatment. For the genomic library, after two rounds, 6 of 192 (3%) phages retained at least 80% of their barstar binding activity. After three rounds, 31 of 86 (36%) phages retained 80% or more of their barstar binding activity. Selection therefore clearly enriched phages displaying protease-resistant p3 fusions.

Example 4

Sequence analysis of selected chimaeric proteins

As an initial characterisation of the selected chimaeric fusion proteins, the sequences of the selected clones from Example 3 were determined. The chimaeric genes of all the 24 most stable phage clones selected from the plasmid-derived library had an open reading frame from the genes for barnase, through the one for chimaeric protein and to the end of the p3 gene. They also contained no stop codons (in addition to the opal stop codon at the 3' end). Twenty of these contained inserts originating from the CspA gene in the correct reading frame. These 20 comprised three different clones (A1 was found 12-times. D6 6-times, G4 twice). Phage A1 contains a deleted version (residues 1 to 52) of the CspA wild type gene, which must have been created through a deletion within a phagemid clone originally harbouring a larger insert (Table 1). Phage D6 contains in addition to the N-terminal half of CspA (residues 1 to 36 as part of the cloning vector) the core of CspA (residues 17 to 53) (Table 1). Phage G4 contains as an insert a partial duplication of the N-terminal half of CspA (residues 2 to 19). Thus from the plasmid-derived library phages with p3-fusion chimeras, in which the N-terminal half of CspA was complemented with another fragment from CspA, were strongly enriched by proteolytic selection.

The sequences of 25 protease resistant phage clones selected from the genomic library revealed 11 different clones (2 clones were found five times, 1 clone four times, 3 clones twice). All inserts kept the reading frame from barnase into p3. They all contained the opal stop codon at their 3'end but no additional stop codons. The inserts of all phages sequenced could be traced back to the *E. coli* genome showing an error-rate of about 1% presumably due to their generation by PCR. 64% of the sequenced phages contained inserts, whose reading frame was identical to that of the originating *E. coli* protein. This suggests an enrichment for DNA fragments in their natural reading frame, as from a random distribution based on three possible reading frames and two possible orientations of any DNA only 16% of inserts would be expected to retain the natural reading frame. However, the selection of clones which originated from open reading frames (ORFs) that do not correspond to the natural reading frame of the originating gene in 36% of the

sequenced inserts indicates that these may also lead to the formation of stably folded chimaeras.

As outlined in Example 1, 60% of all clones in the unselected genomic library contained an opal (TGA) stop codon at the 3' end of the chimaeric gene while the remainder contained the Gly-encoding GGA codon in this position. However, only clones containing opal stop codons at this position were found after proteolytic selection from the genomic library. In the absence of a constitutive stop codon almost exclusively chimaeric gene fusions leading to a frameshift between the barnase and p3 genes were selected (data not shown). These results show that the efficiency (up to 3% according to Eggertsson & Söll, 1988), with which transfer-RNA^{Trp} can decode TGA as a tryptophan, leads to sufficient display of the barnase-chimera-p3 fusion on the phage but appears to reduce folding related, toxic effects. The use of a opal stop codon in the genes encoding the displayed fusion proteins was therefore advantageous for selection in the presented examples.

Example 5

Proteolytic stability of selected chimaera-phages in solution

To show that the sequenced fusion proteins were not only proteolytically stable after immobilisation of the displaying phage on a barstar coated surface (as shown in Example 3) but also in solution, they were tested for proteolytic stability through exposure to trypsin and thermolysin in solution (prior to immobilisation) at different temperatures (Fig. 1a). Phages retaining the barnase tag (as a consequence of a proteolytically stable fusion protein) were captured on barstar and the percentage of retained barstar binding activity was quantitated by ELISA.

Among the phages from the plasmid-derived library two clones (A1 and D6) retained at least 80% of their binding activity after treatment at 20°C. From the genomic library 8 of the 11 clones (1C2, 1G6, 1A7, 2F3, 1B11, 2F1, 2H2, 3A12) retained at least 80% of their activity after trypsin/thermolysin treatment at 24°C. The remaining phages were less well protected from proteolytic attack in solution than when bound to the barstar coated surface (compare Example 3).

Example 6

Soluble expression of selected chimaeric proteins

To characterise the selected chimaeric proteins outside the context of the barnase-p3 fusion protein, the genes of the ten most stable chimaeras of the selected clones in Example 5 were expressed without the fusion partners. For this, their genes were subcloned for cytoplasmic expression into a His-tag vector. Five of these proteins (His-a1, His-d6 from the plasmid-derived library; His-1c2, His-2f3 and His-1b11 from the genomic library) could be purified after expression directly from the soluble fraction of the cytoplasm via their His-tag. The remaining proteins formed inclusion bodies in the expressing cells. One of these, His-1g6 containing an insert expressed in a reading frame different from that of its originating gene (Table II), was refolded via solubilisation in 8M urea. The remaining clones were not further studied.

Example 7

Biophysical characterisations of chimaeric proteins

The first biochemical analysis of the purified chimaeric proteins described in Example 6 concerned their multimerisation status. The chimaeric proteins His-a1, His-d6, His-1c2, His-2f3, His-1g6 formed only monomers according to their elution volume in gel filtration, while His-1b11 formed 30% monomers with the remainder forming dimers.

To analyse the type of secondary structure formed by these chimaeras, the purified proteins were studied by CD and NMR. The CD spectra (Fig. 2a) of the monomeric proteins and the monomeric fraction of His-1b11 were all characteristic of β-structure containing proteins with minima between 215 nm and 225 nm (Greenfield & Fasman 1969, Johnson 1990). All proteins exhibited cooperative folding characteristics with sigmoidal melting curves (Fig. 2b) and midpoints of unfolding transition between 46°C and 62°C (Table I). The cooperative folding behaviour is a strong indication that each of

the analysed chimaeras forms a domain with a single fold, in contrast to a mixture of folded or partially folded structures as in a molten globule.

The NMR spectra of His-2f3 and His-1c2 further suggested the presence of well folded protein domains, as can be inferred from the chemical shift dispersion of many amide protons to values downfield of 9 ppm (Fig. 3a, c) and of methyl group protons to values around 0 ppm in their NMR spectra (Wüthrich 1986). Finally, downfield chemical shifts of C^{α} protons to values between 5 and 6 ppm, as seen in the NMR spectrum of His-1c2 (Fig. 3e), are also frequently observed in β-sheet containing polypeptides like the immunoglobulin domains (Riechmann & Davies 1995).

To determine the thermodynamic stability of the selected chimaeras, the energy of unfolding (ΔG) of the six proteins was inferred from their thermodenaturation curves as measured by CD (Fig. 2b). The folding energies of His-a1, His-d6, His-1b11, His-2f3 and His-1g6 are between 1.6 and 2.4 kcal/mol (Table I). These values are lower than those of typical natural proteins and similar to the so far most stable of the *de novo* designed β-structure proteins, betadoublet (2.5 kcal/mol; Quinn *et al.*1994). However, the His-1c2 protein selected from the genomic library had a considerably higher folding energy of 5.3 kcal/mol, which falls within the normal range of natural proteins (5 to 15 kcal/mol; Pace 1990). His-1c2 is indeed 1.7 kcal/mol more stable than His-CspA.

The relative folding stabilities of His-2f3 and His-1c2 were confirmed through the rate of exchange of their amide protons in D_2O as observed in NMR experiments. For His-2f3 a 1D-^{1}H NMR spectrum recorded after incubation for 24 hours in D_2O buffer at 25°C revealed the complete exchange of its amide protons (Fig. 3a, b). In contrast, amide exchange in His-1c2 was slow allowing the observation of many amide protons in a 1D-^{1}H NMR spectrum after 24 hours at 25°C in D_2O (Fig. 3c,d). A group of amide signals between 8.7 and 10 ppm was even detectable three weeks later at about 40% of their original intensity.

Example 8

Proteolytic stability of chimaeras as soluble proteins

Apart from the spectroscopic evidence for folding stability (see Example 7), stability was also confirmed by the exposure of the isolated chimaeric proteins to proteases in solution. The stability data described in Example 7 of the soluble chimaeric proteins from Example 6 largely correspond to the degree of their protection from proteolysis by trypsin, thermolysin (both used during the selection) and chymotrypsin (Fig.1b). Tryptic degradation of the N-terminal His-tag through cleavage after Arg11 was observed for all six proteins. This arginine was introduced as part of the expression vector immediately C-terminal of the N-terminal His-tag. His-1c2 (with a folding energy of 5.3 kcal/mol) is no further degraded by any of the proteases confirming its high conformational stability, but the other proteins are partially proteolysed within the main body of the polypeptides. This is consistent with a partial unfolding expected from a folding energy of about 2 kcal/mol. Thus although all the proteins are resistant to proteolysis (for example compared with the facile cleavage of the His-tag at Arg), the resistance varies between the proteins and upon the conditions.

Example 9

Sequence duplications in selected chimaeric proteins

As outlined in Example 4 above, in 20 of 24 sequenced chimaeric proteins, which were selected from the plasmid-derived library, the N-terminal half of CspA was complemented with another fragment from CspA. Indeed, the chimaeric proteins D6 and G4 both comprise a partial duplication of their N-terminal half. Phage D6 contains in addition to the N-terminal half of CspA (residues 1 to 36 as part of the cloning vector) the core of CspA (residues 17 to 53) (Table 1). Phage G4 contains as an insert a partial duplication of the N-terminal half of CspA (residues 2 to 19). This result indicates that (partial) duplication of amino acid segments can lead to the formation of stably folded protein domains.

Example 10

Duplication of homologous elements in stably folded chimaeric proteins

No direct structural information is available for the seven DNA fragments. which were found after selection of the genomic library (Example 1) and which were expressed in their natural reading frame. One however has a high level of sequence identity with a sequence neighbour of known three dimensional structure (as identified by BLAST analysis of the *E. coli* genome). The insert of phage 1B11 spans residues 364 to 398 in the *E. coli* 30S ribosomal subunit protein S1 (gene identifier 1787140), of which residues 369 to 397 have a 52% identity with residues 11 to 39 of S1 RNA-binding domain from the *E. coli* polynucleotide phosphorylase. These comprise a stretch of four β-strands in the 3D structure of the S1 domain, which like CspA forms a β-barrel albeit with an inserted helix (Bycroft *et al.* 1997).

The two S1 domains (of the 30S ribosomal protein and of the phosphorylase) are according to their sequence similarity and identity homologous to CspA. The juxtaposition of the segments in the chimaeric protein 1B11 represents therefore a juxtaposition of corresponding regions from homologous polypeptide domains (which also forming the same structural fold). This result indicates that a (partial) duplication of homologous amino acid segments can lead to the formation of stably folded protein domains.

Example 11

Evidence for complementation with elements of similar structure from proteomic analysis in a chimaeric protein

20 of the 24 most stable phage clones selected from the plasmid-derived library (Example 2) contained inserts originating from the CspA gene in the correct reading frame (see Example 4). These 20 comprised three different clones (A1, D6, G4). A1 contains a deleted version (residues 1 to 52) of the CspA wild type gene, which must have been created through a deletion within a phagemid clone originally harbouring a larger insert

(Table 1). Phage D6 contains in addition to the N-terminal half of CspA (residues 1 to 36 as part of the cloning vector) the core of CspA (residues 17 to 53) (Table 1). Phage G4 contains as an insert a partial duplication of the N-terminal half of CspA (residues 2 to 19). The complementing sequences in all three clones comprise regions of CspA, which in the CspA structure form β-strand regions. Thus sequences forming the same type of secondary structure are juxtaposed in the chimaeric proteins A1, D6 and G4.

No direct structural information is available for the seven DNA fragments, which were found after selection of the genomic library (Example 1) and which were expressed in their natural reading frame. One however has a high level of sequence identity with a sequence neighbour of known three dimensional structure (as identified by BLAST analysis of the *E. coli* genome). The insert of phage 1B11 spans residues 364 to 398 in the *E. coli* 30S ribosomal subunit protein S1 (gene identifier 1787140), of which residues 369 to 397 have a 52% identity with residues 11 to 39 of S1 RNA-binding domain from the *E. coli* polynucleotide phosphorylase. These comprise a stretch of four β-strands in the 3D structure of the S1 domain, which like CspA forms a β-barrel albeit with an inserted helix (Bycroft *et al.*1997). Thus sequences forming the same type of secondary structure are juxtaposed in the chimaeric protein 1B11.

Thus in case of the His-a1, His-d6 and His-1b11 proteins the juxtaposition of sequences, which form the same type of secondary structure, have lead to the formation of stably folded chimaeric protein. Overall, gene fragments selected from both libraries appear to be enriched for sequences forming primarily β-structure in their parent protein. Such sequences may be more frequently able to from a stable domain with another gene fragment, that originally encodes part of a β-barrel, than sequences of a helical origin.

Example 12

Evidence for complementation with elements of different structure from proteomic analysis in selected chimaeric proteins

No direct structural information is available for the seven DNA fragments, which were found after selection of the genomic library (Example 1) and which were expressed in

their natural reading frame. One however has a high level of sequence identity with a sequence neighbour of known three dimensional structure (as identified by BLAST analysis of the *E. coli* genome). The insert of 3A12 spans residues 52 to 80 in the putative transport periplasmic protein (gene identifier 1787590) sharing a 48% sequence identity with residues 30 to 58 of the Salmonella oligopeptide-binding protein. In its 3D structure (Tame *et al.*1994) these residues form a helix and two short antiparallel β-strands. The oligopeptide-binding protein as a mixed α/β protein has no structural homology with CspA and its residues 52 to 80 do not form part of a β-barrel. Thus sequences from different folds are juxtaposed in the chimaeric protein 3A12. Thus while gene fragments selected from both libraries appear to be enriched for sequences forming primarily β-structure in their parent protein, polypeptide sequences originating from different folds are also represented.

Example 13

Effects of modified selection conditions

Proteolytic selection seemingly favoured phages displaying chimaeric proteins with higher folding stabilities than those displaying chimeras with high melting points. From the plasmid-derived library the phage clone displaying the more stable protein A1 was selected twice as frequently as the less stable D6, which however has the higher melting point (Table I). In case of the genomic library the phages displaying the two most stable proteins (1C2, 1G6) were found four and five times, while the phages of the two less stable proteins (1B11, 2F3) were only found twice each after selection. Again the His-1b11 and His-2f3 proteins have the higher melting points (Table I). This suggests that escape from proteolysis depends more on stability than on the melting point as long as proteolysis is performed at temperatures well below melting points. Higher proteolysis temperatures than used here may therefore allow more frequent selection for proteins with higher melting points, while energetically more stable proteins would probably be enriched if phages are proteolysed for longer.

Such modified conditions may increase the frequency, with which polypeptides exhibiting stabilities of natural proteins are selected from random combinatorial libraries. Further

improvements may be expected by use of much larger repertoires, for example created by scale up, by improvements in the transfection efficiency of plasmid, phagemid or phage replicons into cells, or by other techniques such as *in vivo* recombination using the cre-lox system (Sternberg & Hamilton 1981). Alternatively or in addition repertoires could be further diversified by mutagenesis before or after selection. Effective repertoire sizes can further be increased, when recombination partners are enriched prior to recombination for in frame, no stop codon containing DNA fragments.

The presented methodology allows the selection of new chimaeric proteins, which have been created through recombination of natural genes and which can combine properties from different molecules. Using suitable combinatorial partners polypeptides may be created, which inherit desirable functions (such as a target binding sites or an antigenic epitope) from parent proteins, while removing undesirable properties (such as such as unwanted receptor binding sites or unwanted epitopes). For this purpose, proteolytic treatment may be combined with selection for binding.

In the case of selection for binding of chimaeric proteins to a ligand, it may be advantageous to increase the copy number of phage displayed fusion proteins. An increased copy number of displayed p3-fusion proteins, of which there can be up to five on each phage particle, would result in multiple binding events for a single clone, which may allow selection even in the case of chimaeric proteins with a low affinity to the ligand. Copy number of fusion proteins in phage display can for example be increased, when phagemid-encoded fusion p3-fusion protein are rescued for phage preparation with a helper phage lacking the gene for p3 (Rakonjac *et al.* 1997)

Example 14

Secondary modifications of selected chimaeras

The binding activity of chimaeric proteins created through the random recombination of polypeptide segments for a given ligand may be low, even if the parent proteins of these segments have a high affinity for such a ligand. Thus any newly juxtaposed polypeptide segment is expected to have some effect on the structure of the other when compared with

its structure in the parent protein. As a consequence most binding sites will no longer fit a ligand with the same precision and result in a reduced affinity. It is therefore envisaged that it may be necessary to improve such binding sites, once a new chimaeric protein has been created as part of a combinatorial library.

Improvements of selected chimaeric proteins can be achieved by secondary modification or mutation. Such modifications can be made to improve binding, they may also be made to increase stability and/or to introduce new binding or enzymatic functions. The type of modification and its location in the chimaeric protein (i.e. which old amino acid is replaced with which new one) may be based on rational design principles or partially or entirely random. Modifications can be introduced by a site-directed mutagenesis (Hutchison III *et al.*1978) or by a site-directed random mutagenesis (Riechmann & Weill 1993) followed by selection or screening for activity or stability in the resulting mutant chimaeras. Alternatively an entirely random mutagenesis (through for example error-prone PCR amplification, Hawkins *et al.*1992) of either one or both segments (or indeed their linking sequence) of the chimaeric protein or through passage of the phagemid through an *E. coli* mutator strain (Low *et al.*1996) followed by selection and/or screening for binding, enzymatic activity or stability.

Modifications can further comprise the deletion of residues or introduction of additional residues. In particular the joining and end regions of the recombined polypeptide segments may be expected to be not optimised. The joining regions may strain interactions between the juxtaposed segments, which may be relieved by introducing additional residues within the joining region. Regions close to the end of the chimaeric protein may comprise terminal residues not participating in the fold of the domain, and their deletion may improve the overall integrity of the protein.

We demonstrate for one of the chimaeric protein how its stability was improved based on rational design. His-2f3 was created through the combinatorial shuffling of the N-terminal half of the *E. coli* protein CspA with random amino acid segments encoded by fragments of the *E. coli* genome (Example 1). The sequence and genetic origin of the random fragment are given in Table II. The spectroscopic analysis of His-2f3 (Example 7)

indicates a fold rich in β-structure. If His-2f3 folds (like CspA) into a β-barrel certain sequence requirements may have to be met to improve the stability of the barrel.

In CspA the hydrophobic side chain of residue Leu45 closes one end of its β-barrel and Gly48 and Gln49 form a turn between two β-strands in the polypeptide fold to allow the formation of backbone hydrogen bonds of the following β-strand with the N-terminal β-strand of CspA. Within this strand the side chains of three hydrophobic residues (Val51, Phe53 and Ile55) point to the inside of the barrel. His-2f3 does not meet those requirements exactly but has a similar motifs within its genomic segments, as the residues Pro58, Gly61, Ala62, Met64, Phe66 and Ala68 (in its genomic segment) exhibit the same spacing as the motif described for CspA (compare Table III).

We therefore mutated the genomic segment in 2f3 at positions 58 (P to L) , 62 (A to Q) and 68 (A to L) to match the amino acid types described for the motif in CspA, while the residues at position 61, 64 and 66 in 2f3 were already judged to be identical or similar enough. As summarised in Table III, the combined P58L and the A62Q mutations increased the stability of 2f3 to 6 kcal/mol, which lies within the range of typical natural protein domains and is 1.6 kcal/mol higher than that of CspA itself. The A68L had no positive effect in 2f3.

In addition, the two C-terminal residues (PW) in 2f3 (compare Tables II and III) were removed, which were partially degraded in the originally expressed, soluble 2f3 protein. The removal of these residues had no significant effect on the overall stability of 2f3, but resulted in a more homogeneous protein preparation after expression, which for example is advantageous for structural studies like NMR.

This result shows that new chimaeric proteins can be improved upon after selection through further modifications, in this case based on rational design.

Example 15

Crossreactivity of anti-CspA antisera with chimaeric proteins

A possible application of chimaeric proteins is their use as vaccines against the parent polypeptide of one or more of the recombined amino acid sequences. For this purpose antisera against the chimaeric protein will be cross-reactive with the parent polypeptide (and indeed vice versa).

A rabbit was immunised with CspA using Freunds adjuvant (see Methods). The resulting antiserum recognised immobilised, biotinylated CspA. Binding of the rabbit antiserum to the immobilised Biotin-CspA could be competed with soluble CspA and to varying degrees with the chimaeric protein His-1c2, His-2f3 and His-1b11 (Fig. 4). This result shows that an immunisation with CspA results in an immune response which contains antibodies that crossreact with all three of the analysed chimaeras. Conversely, it should therefore also be possible to achieve an immune response against CspA when any one of the chimaeras is used for vaccination. The immune response can be expected to be directed against both linear and against conformational determinants of CspA.

Example 16

Selection of chimaeric proteins for binding

In the earlier examples, stably folded chimaeric domains were selected by proteolysis through the combinatorial juxtaposition of the N-terminal half of the *E. coli* protein CspA with amino acid segments encoded by fragments of the *E. coli* genome (Examples 1 and 3). A number of these chimaeric proteins are expected to form a polypeptide fold resembling that of CspA as the secondary structure prediction and spectroscopic analyses of the four chimaeras described (Example 7) indicates a fold rich in β-structure.

It is possible that the RNA binding function (Jiang *et al.*1997) of CspA is retained in some of the selected chimaeras. The nucleic acid binding site in CspA has been proposed

to be located on a surface formed around Trp11, Phe18, Phe20, Phe31 and Lys60 (Newkirk *et al.*1994; Schroder *et al.*1995). While the four aromatic residues are part of the N-terminal half of CspA and are therefore present in all members of the genomic repertoire (Example 1), residue Lys60 is not. It seems likely that in some of the chimaeric proteins the nucleic acid binding activity will be retained; such proteins could be selected for example by binding of phage displaying the protein to nucleic acid immobilised on solid phase. (However as the phage display system used in the experiments above would be unsuitable as the barnase tag retains nucleic acid binding activity).

Furthermore, for functional selection it may be necessary to use a phage-display system which allows the multiple display of the fusion protein thereby facilitating selection of chimaeric proteins with low affinities for the ligand (in this case nucleic acid) through the resulting avidity effect. This may be achieved in the case of chimaeras fused to the phage coat protein p3 for example through the use of a phage vector like phage fd (Zacher *et al.*1980), through the use of a phagemid in combination with a helper phage devoid of the phage p3 gene (Rakonjac *et al.*1997) or through an increased expression of functional chimaera-p3-fusion protein. Alternatively, multiple display may be achieved through fusion to a different phage coat protein, like p8.

Of particular importance is the binding of the chimaeric domains to antibodies. If antiserum against the parent protein were used for selections, this would be expected to direct the selection to any of the epitopes of the chimaeric protein that are similar to those in the parent protein and are represented in the anti-serum. Alternatively monoclonal antibodies could be used which would select for those clones binding a single epitope that is similar to that of the parent protein. A number of these chimaeric proteins are expected to form a polypeptide fold resembling that of CspA, as the secondary structure prediction and spectroscopic analyses of the four chimaeras described in Example 7 indicates a fold rich in β-structure. If any of the recombined chimaeric proteins within the repertoire resemble in fold that of CspA, it should therefore be possible to enrich for such proteins through binding to antibodies which specifically recognise CspA.

Example 15 already describes that an anti-CspA antiserum crossreacts with three of the chimaeric proteins selected through proteolysis (and barstar binding) alone. The anti-

CspA antiserum may therefore serve as a reagent to enrich the combinatorial library from Example 1 specifically for phages displaying chimaeric proteins which resemble CspA most closely.

A rabbit anti-CspA serum was fractionated through binding to Biotin-CspA immobilised to Streptavidin-agarose to enrich for that against conformational determinants of CspA. Purified CspA-specific (anti-CspA) rabbit antibodies (IgG) were tested for anti-CspA binding activity as described in Example 15. For use in phage selection the anti-CspA rabbit antibodies were immobilised on a Streptavidin-coated ELISA-well plate through a commercial biotinylated goat anti-rabbit IgG antiserum. Phages ($7x10^9$ cfu) from the genomic library of chimaeric proteins (Example 1), which had undergone one round of proteolytic selection (followed by barstar binding, see Example 3), were treated with trypsin and thermolysin (see Example 3) followed by binding to the CspA-specific rabbit antibodies in 2% BSA in PBS. After washing with PBS and 40 mM DTT $4.3x10^3$ bound phage were eluted at pH2, neutralised and used for infection of bacterial cells.

96 of the resulting clones were grown up in a multiwell plate and infected with helper phage KM13 for phage production. Phage from the culture supernatants from the infected bacterial clones were bound to the anti-CspA antibodies, which again had been immobilised to a Streptavidin-coated plated via a biotinylated goat anti-rabbit IgG antiserum. Bound phage was washed with PBS, exposed to trypsin and thermolysin after immobilisation as before, washed with PBS and remaining phage was detected with an anti-M13-HRP conjugate. Sequences of the nine clones with the strongest signal remaining after proteolysis were determined. Seven of these clones were identical (two) or almost identical (five clones had one residue less at the N-terminal end of the genomic insert and two different residues at the C-terminal end) to the clone 2f3, which had been previously selected - albeit not at the same high frequency - after proteolytic selection/barstar binding (Examples 3 and 4). The two remaining sequences had not been previously observed. Purified phage of the 2f3 and 2f3-like clones was confirmed to be strongly reactive with the purified rabbit anti-CspA antibodies, also after exposure to trypsin in solution, confirming that it is protease-resistant folded sequences that are binding to antibody. Together with the fact that the anti-serum had been fractionated for binding to the folded CspA, this indicates strongly that the selection has been towards a

conformational determinant. The ELISA in Fig. 4 (see Example 15) proves that the corresponding chimaeric protein also interacts in its soluble version with an anti-CspA antiserum.

This experiment suggest how it is possible to identify "isosteric" peptides (same conformation in parent protein and chimaeric domain). It also indicates that the method can be used for vaccination towards a conformational segment of the protein; thus it should equally be possible to use 2f3 for vaccination and to produce anti-serum that recognises the conformation of the N-terminal portion of CspA.

Methods

Vector constructions

The gene for the H102A mutant of barnase (Meiering *et al.*1992) was fused to the N-terminus of the gene 3 protein (p3) of phage fd (Zacher *et al.*1980) in a modified phagemid pHEN1 (Hoogenboom *et al.*1991) between the DNA encoding the pelB leader peptide and the mature p3 after PCR amplification with suitable oligonucleotides using NcoI and PstI restriction sites to create the vector p22-12. Into p22-12 suitably amplified parts of the *E. coli* gene CspA (Goldstein *et al.*1990) were cloned between the barnase and the p3 genes using PstI and NotI restriction sites. In the resulting phagemid vector pC5-7 the barnase gene is followed by the N-terminal 36 residues of CspA (the N-terminal Met being mutated to Leu to accommodate the PstI site) and the DNA sequence GGG *AGC TCA GGC GGC CGC* AGA A (SacI and NotI restriction sites in italics) before the GAA codon for the first residue (Glu) of p3. In pC5-7, the barnase-Csp cassette is out of frame with the p3 gene. In the control vector pCsp/2 the barnase-Csp cassette is in frame with the p3 gene, but the first codon of the linking DNA constitutes an opal stop codon.

Vectors for the cytoplasmic expression of soluble proteins were constructed by subcloning of genes from the phagemids into the BamHI and HindIII sites of a modified QE30 vector (Qiagen). This vector is identical to QE30 except for a tetra-His tag. During PCR-aided subcloning using the primers CYTOFOR (5'-CAA CAG TTT *AAG CTT* CCG CCT GAG CCC AGG-3') and CYTOBAK (5'-CCT TTA CA*G GAT CC*A GAC TGC AG-3') opal stop-codons were converted into the Trp-encoding TGG triplet.

Library construction

As templates for random amplifications 100 ng of a pBCSK (Stratagene) based plasmid containing the entire CspA coding region or genomic DNA (2 µg digested with SacI) from the *E. coli* strain TG1 (Gibson 1984) prepared as described (Ausubel *et al.*1995) was used in 25 PCR cycles with an annealing temperature of 38°C using the oligonucleotide SN6NEW (5'-GAG CCT GCA *GAG CTC* AGG NNN NNN-3' at 40 pmole/ml) for the plasmid or in 30 PCR cycles with an annealing temperature of 38°C using the oligonucleotide SN6MIX (5'-GAG CCT GCA *GAG CTC* CGG NNN NNN-3' at 40 pmole/ml) for the genomic DNA. PCR products were extended in a further 30 cycles with an annealing temperature of 52°C using the oligonucleotide NOARG (5'-CGT GCG AGC CTG CA*G AGC TCA* GG-3' at 4.000 pmole/ml) for the plasmid and the oligonucleotide XTND (5'-CGT GCG AGC CTG CA*G AGC TCC* GG-3' at 4.000 pmole/ml) for the genomic DNA. PCR products of around 140 bp were purified from an agarose gel and reamplified in 30 PCR cycles using the oligonucleotide NOARG at an annealing temperature of 50°C.

Resulting fragments were digested with SacI, purified and ligated into the phosphatased and SacI-digested vector pC5-7. Ligated DNA was electroporated into TG1 creating a plasmid-derived repertoire of $1.7x10^8$ clones and a genomic repertoire of $1.0x10^8$ clones. In both libraries about 60% of the recombinants contained monomeric inserts, while the remainder contained oligomeric inserts. Ligation background was less than 1% for both ligations. Due to differences in the 3' end of the PCR primers XTND and NOARG 40% of clones with in-frame inserts in the genomic library contained a GGA-encoded Gly residue as part of the 3'-SacI site, while the remaining clones contained the TGA-encoded opal stop-codon at the same position. All members of the plasmid-derived library with in-frame inserts contained the TGA-encoded opal stop-codon at this position.

Selections

For selections about 10^{10} colony forming units (cfu) of phage were treated with 200 nM trypsin (Sigma T8802) and 384 nM thermolysin (Sigma P1512) in TBS-Ca buffer (25 mM Tris, 137 mM NaCl, 1mM $CaCl_2$, pH 7.4) for 10 minutes at 10°C. Proteolysed phage was captured for 1 hour with biotinylated C40A,C82A double mutant barnase inhibitor barstar (Hartley 1993, Lubienski *et al.*1993) immobilised on a streptavidin

coated microtitre plate (Boehringer) wells in 3% Marvel in PBS. Wells were washed twenty times with PBS and once with 50 mM dithiothreitol (DTT) in PBS for 5 minutes to elute phage containing proteolysed p3-fusions held together solely by disulphide bridges. Bound phage was eluted at pH 2, neutralised to pH 7 and propagated after reinfection.

Phage ELISA

Proteolysis and binding of purified phage (about 10^{10} cfu per well) to immobilised barstar was performed as above. Phage remaining bound after washes with PBS and DTT was detected in ELISA with an anti-M13 phage antibody - horse radish peroxidase (HRP) conjugate (Pharmacia) in 3% Marvel in PBS. Non-purified phage from culture supernatants was bound to the biotinylated barstar and then proteolysed *in situ*. Purified phage was proteolysed in solution and proteases were inactivated with Pefabloc (Boehringer) and EDTA before capture.

Anti-CpsA antisera

A first anti-CspA serum (as used for Fig. 4) was obtained from an immunised rabbit. The rabbit was injected once with refolded (see above) His-CspA (0.5 ml at 1.75mg/ml PBS) mixed with 1:1 with Freud's complete adjuvant, followed by two injections with refolded His-CspA (0.5 ml at 1.75mg/ml PBS) mixed 1:1 with Freud's incomplete adjuvant in 4 week intervals to boost the immune response. The antisera used was obtained from blood taken ten days after the second boost.

A second anti-CspA serum (as used for purification of anti-CspA specific antibodies in Example 16) was obtained from a different immunised rabbit. The rabbit was injected once with refolded (see above) His-CspA (0.5 ml at 1.75mg/ml PBS) mixed with 1:1 with Freud's complete adjuvant, followed by three injections with refolded His-CspA (0.5 ml at 1.75mg/ml PBS) alone in 4 week intervals to boost the immune response. The antisera used was obtained from blood taken ten days after the third boost. One ml of this antiserum was purified on 0.2 ml of Streptavidin-agarose (Pierce No. 53117), to which about 0.1 mg Biotin-CspA (see below) was bound, after washing with PBS, elution at pH 2 followed by neutralisation and buffer exchange into 3.5 ml PBS (i.e. at 3.5 fold dilution compared to the original antiserum). When used for phage selection, the purified anti-

CspA antibodies were 500-fold diluted in PBS for binding to a biotinylated goat anti-rabbit antiserum (Sigma B-7389) immobilised in Streptavidin-coated ELISA wells.

His-CspA, as used for immunisation and data in Table III and Fig. 4, was purified from the unfractionated *E. coli* cell pellet using NTA agarose after solubilisation with 8M urea in TBS. Before elution with 200 mM imidazole in PBS, agarose bound His-CspA was renatured with an 8M to 0M urea gradient TBS. Eluted protein was dialysed against PBS.

For biotinylation, CspA was modified through addition of cysteine-glutamine-alanine residues as a C-terminal tag, introduced on the gene level using suitable PCR primers. The corresponding His-CspA-Cys protein was expressed, purified and refolded as His-CspA except for the addition of 0.5 mM DTT to all solutions. The NTA agarose with the bound His-CspA-Cys was washed with 5 volumes of PBS (all solutions without DTT from this step onwards) and mixed with the biotinylation reagent EZ-Link™ Biotin-HPDP (Pierce) for biotinylation according to the manufacturer's instructions. After 1 hour the agarose with the bound and biotinylated protein was washed with 10 volumes of PBS, eluted with 200 mM imidazole in PBS and buffer-exchanged into PBS. Biotinylation of the now His-Biotin-CspA was verified by MALDI mass spectrometry using a SELDI (Ciphergen systems).

Binding of rabbit anti-CspA antisera to CspA was analysed after immobilisation of biotinylated His-Csp-Cys (at 0.25 µg/ml in PBS) onto streptavidin-conjugated ELISA plates (Boehringer Mannheim). The rabbit anti-CspA serum was diluted 1/30,000 in 2% bovine serum albumin in PBS and preincubated with varied amounts of purified competitors (see Fig. 4) before binding to the ELISA well. Bound rabbit antibodies from the serum were detected with a HRP-conjugated goat anti-rabbit IgG antiserum (Sigma).

2f3 mutants

The gene for the 6H-2f3 protein (compare Table III) was prepared by PCR with the primers QEBACK (5'-CGG ATA ACA ATT TCA CAC AG-3') and 2F3FOR (5'-GGC CGC CTG *AAG CTT* TTA AGG CGG ATG GTT GAA-3') using the 2f3 gene in QE30 (compare Table II) as a template. Mutant genes for the 6H-2f3 protein were prepared through PCR amplification of the partial 2f3 gene using accordingly designed primers and

the same template. For each mutant two PCR products (covering the N and C-terminal portion of the 2f3 gene respectively) were purified, denatured, annealed and extended. Full-length mutant genes were specifically reamplified using the two outside primers BACKTWO (5'-CCT TTA CAG GAT CC-3') and 2F3FOR. Complete genes were digested with HindIII and BamHI and cloned into the unmodified QE30 vector (Qiagen; encoding a 6 histidine containing N-terminal tag).

For the mutant 6H-2f3-P58L the primers 2F3F2 (5'-GGT AAA AAG CAT GAT TGC GCC AAT TTC TAG CTC GCC TGC-3'), CYTOBAK (for the N-terminal half), 2F3B0 (5'-GGT AAA AAG CAT GAT TGC G-3') and QEFOR (5'- GTT CTG AGG TCA TTA CTG G-3') (for the C-terminal half were used). For the mutant 6H-2f3-P58L,A62Q the primers 2F3F1 (5'-GGT AAA AAG CAT GAT TTG GCC AAT TTC TAG CTC GCC TGC-3'), CYTOBAK (for the N-terminal half), 2F3B0 and QEFOR (for the C-terminal half were used). For the mutant 6H-2f3-P58L,A62Q,A68L the primers 2F3F1, CYTOBAK (for the N-terminal half), 2F3B1 (5'-AAT CAT GCT TTT TAC CCT AAT GGA TGG C-3') and QEFOR (for the C-terminal half were used).

Protein expression, purification and analysis

Proteins were expressed by induction of exponential bacterial cultures at 30°C and purified from the soluble fraction of the cytoplasm using NTA agarose according to the Qiagen protocol. His-1g6 was purified after solubilisation with 8 M urea in TBS and refolded by dialysis from 8 M, 4 M, 2 M, 1 M, 0.5 M to 0 M urea in TBS. Proteins were further purified by gel filtration on a Superdex-75 column (Pharmacia). The molecular weight of proteolytic fragments was determined using the surface enhanced laser desorption/ionisation (SELDI) technique (Hutchens & Yip 1993).

Proteolysis of soluble proteins (about 40 μM) was carried out using 40 nM of trypsin, thermolysin or α-chymotrypsin (Sigma C3142) in TBS-Ca at 20°C for 10 minutes. Circular dichroism spectra and thermodenaturation were recorded as described (Davies & Riechmann 1995). Thermodenaturation of 10 μM protein (His-1c2 at 2 μM) in PBS was followed at a wavelength between 220 nm and 225 nm (His-1c2 in 2.5 mM phosphate buffer, pH 7, at 205 nm). Nuclear magnetic resonance experiments were performed on a Bruker DMX-600 spectrometer as described (Riechmann & Holliger 1997) using a

watergate sequence (Piotto *et al.* 1992) for water suppression with protein at 1 mM in 20 mM phosphate buffer at pH 6.2 containing 100 mM NaCl in 93% H_2O / 7% D_2O or 99.9% D_2O.

All publications mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described methods and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed various modifications of the described modes for carrying out the invention which are obvious to those skilled in molecular biology or related fields are intended to be within the scope of the following claims.

Table I. Amino acid sequences and biophysical parameters of de novo proteins.

Protein	Tm,°C	ΔG[1]	MW, Da	C-terminal sequence[2]
His-Csp	59.8	3.6	8.565	IQNDGYKSLDEGQKVSFTIESGAKGPAAGN VTSLEA
His-Csp/2	no expr.[3]		5,854	WAQAEA
plasmid library:				
His-a1	46.0	2.1	7,729	IQNDGYKSLDEGQKVSFTWAQAEA
His-d6	47.6	1.6	10,352	GSSGFGFITPDDGSKDVFVHFSAIQNDGYK SLDEGQKVSFTWAQAEA
genomic library:				
His-1b11	57.1	2.0	10,722	GSSGAAVRGNPQQGDRVEGKIKSITDFGIF IGLDGGIDGLVHLSDISWAQAEA
His-2f3	61.4	1.8	10,582	GSSGAGEPEIGAIMLFTAMDGSEMPGVIRE INGDSITVDFNHPPPWAQAEA
His-1c2	54.8	5.3	10,972	GSSGRVISLTNENGSHSVFSYDALDRLVQQ GGFDGRTQRYHYDLTWAQAEA
His-1g6	48.4	2.4	10,485	GSSGKSGVKTDYRASASIACAYAGAGSSDS RRSFLCITRSESDGPWAQAEA

1: The conformational stability ΔG (kcal/mol) at a given temperature T was calculated using the Gibbs-Helmholtz equation $\Delta G(T) = \Delta Hm\ (1-T/Tm) - \Delta Cp\ [(Tm-T) + \ln(T/Tm)]$, while inferring the midpoint of thermal unfolding (Tm) and the enthalpy change for unfolding (ΔHm) at the Tm from the denaturation curve (Agashe & Udgaonkar 1995) and assuming for ΔCp (the difference in heat capacity between unfolded and folded conformation) a value of 12 cal per residue (Edelhoch & Osborne 1976).

2: Sequences shown are those following the N-terminal half of CspA, which is MRGSHHHHGSRLQSGKMTGIVKWFNADKGFGFITPDDGSKDVFVHFSA.

3: The His-(Csp/2) protein was found neither in the soluble nor insoluble fraction of the cytoplasm presumably due to degradation within the cell.

Table II. Sequences and origin of genomic segments

Segment[a]	Sequence[d]	Genetic origin[e]	Protein origin[f]
1a7[b]	GIATSAICDA QVIGEEPGQP TSTTCRFRSK FSAIAFPW	8931 to 9041 in ECAE298, gatC	minus strand
1b11[b,c]	GAAVRGNPQQ *GDRVEGKIKS ITDFGIFIGL DGGIDGLVHL SDISW*[g]	6382 to 6514 in ECAE193, rpsA	364 to 398 in RS1_ECOLI
1c2[b,c]	*GRVISLTNEN GSHSVFSYDA LDRLVQQGGF DGRTQRYHYD LTW*	2178 to 2303 in ECAE156, rhsD	645 to 686 in RHSD_ECOLI
1g6[b,c]	GKSGVKTDYR ASASIACAYA GAGSSDSRRS FLCITRSESD GPW	2694 to 2569 in ECAE116, rluA	frameshift
2f1[b]	*GAGTMAEEST DFPGVSRPQD MGGLGFWYRW NLGWMHDTLD YMK*PHSW	8558 to 8422 in ECAE419, glgB	452 to 494 in GLGB_ECOLI
2f3[b,c]	*GAGEPEIGAI MLFTAMDGSE MPGVIREING DSITVDFNHP* PPW	5431 to 5551 in ECAE113, slpA	89 to 127 in FKBX_ECOLI
2h2[b]	GSAYNTNGLV QGDKYQIIGF PRFNQLTVYF HNLPW	7955 to 7854 in ECAE475, yjbC	minus strand
3a12[b]	*GKAVGLPEIQ VIRDLFEGLV NQNEKGEIVP* W	1479 to 1568 in ECAE231, b1329	52 to 80 in MPPA_ECOLI

1g7	GWLKRKLNLK FNEASIAGCD ALLNAAW	ECAE217. b1191	7290 to 7213 in	frameshift
1h12	GC*VPYTNFSL IYEGKCGMSG GRVEGKVIYE TQSTHK*HSW	ECAE485. cadA	12035 to 11927 in	334 to 367 in DCLY_ECOLI
2e2	*GMWPLDMVNA IESGIGGTLG FLAAVIGPGT ILGKIMEVSW*	ECAE324, dsdX	7398 to 7514 in	45 to 83 in DSDX_ECOLI

Segments retaining 80% barstar binding activity after proteolysis of phage *in situ*[a] and in solution[b] and those purified as chimaeric proteins[c]. The sequence of the genomic segment[d] as a C-terminal appendage to the N-terminal region of CspA (LQSGKMTGIV KWFNADKGFG FITPDDGSKD VFVHFSAGSS) is listed; sequences expressed in-frame with the originating gene are shown in italics. The location of each segment within the *E. coli* genome is indicated by nucleotide numbers in the EMBL database entry and name of the originating gene[e], and for those expressed in the same frame of the originating gene, the residue numbers of the corresponding protein and its ID in the Swiss protein database are given[f]. A single base pair deletion after the first 29 base pairs in the DNA insert of 1b11 renders the first 10 residues out of frame with the rspA gene[g].

Table III.

(a) Amino acid sequences of CspA and His-2f3

```
                10          20          30          40          50
CspA MSGKMTGIVK WFNADKGFGF ITPDDGSKDV FVHFSAIQND GYKSLDEGQK
                                                    *  **

2f3  ...SGKMTGIVK WFNADKGFGF ITPDDGSKDV FVHFSAGSSG AGE-PEIGAI
                24          34          44          54          63

                60          70
CspA VSFTIESGAK GPAAGNVTSL
     * * *
2f3  MLFTAMDGSE MPGVIREING DSITVDFNHP P
                73          83          93
```

(b) Folding energy of 2f3 mutants and CspA

Protein	ΔG at 298K (kcal/mol)
CspA	3.4
6H-2f3	1.9
6H-2f3-P58L	2.8
6H-2f3-P58L, A62Q	6.0
6H-2f3-P58L, A62Q, A68L	3.2

(a) The amino acid sequence of CspA is that from the native gene as in the EMPL database. The numbering of the 2f3 sequence takes into account the N-terminal His-tag (MRGSHHHHHHGSRLQ). The C-terminal residues PWAQAEA (compare 2f3 in Table I) were deleted in the constructs used for the data here, as they were partially cleaved in the expressed protein of the original His-2f3 construct indicating that they did not participate to the fold of the chimaeric domain. Their deletion had no significant effect on the overall folding stability of the domain (1.8 vs. 1.9 kcal/mol in the 2f3 constructs used for data in

Table I and III respectively). The residues important for the β barrel fold in CspA as discussed in Example 14 are indicated by an asterisks.

(b) Folding energies were determined as described in Table I. Mutation for 2f3 denote the original amino acid, followed by the residue number and the new amino acid.

References

Ausubel. F.M.. Brent. R.. Kingston. R.E.. Moore. D.D.. Seidman. J.G.. Smith. J.A. & Struhl, K. (1995) Current protcols in molecular biology. Chapter 2.4.1. Wiley & Sons.

Agashe, V.R. & Udgaonkar, J.B. (1995) Thermodynamics of denaturation of barstar: evidence for cold denaturation and evaluation of the interaction with guanidine hydrochloride. Biochemistry 34, 3286-3299.

Alba, E. de, Santoro, J., Rico, M. & Jimenez, M.A. (1999) De novo design of a monomeric three-stranded anti-parallel β-sheet. Protein Sci. 8, 854-865.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller. W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Barbas, C.F., Crowe, J.E., Cababa, D., Jones, T.M., Zebedee, S.L., Murphy, B.R., Chanock, R.M. & Burton D.R. (1992) Human monoclonal fab fragments derived from a combinatorial library bind to respiratory syncytial virus-f glycoprotein and neutralize infectivity. Proc. Natl. Acad. Sci. USA 89, 10164-10168.

Bogarad, L., & Deem, M. (1999) A hierarchical approach to protein molecular evolution. Proc. Natl. Acad. Sci. USA 96, 2591-2595.

Breitling, F., Dubel, S., Seehaus, T., Klewinghaus, I. & Little M. (1991) A surface expression vector for antibody screening. Gene 104, 147-153.

Burton, D.R., Barbas, C.F., Persson, M.A.A., Koenig, S., Chanock, R.M. & Lerner. R.A. (1991) A large array of human monoclonal-antibodies to type-1 human-immunodeficiency-virus from combinatorial libraries of asymptomatic seropositive individuals. Proc. Natl. Acad. Sci. USA 88, 10134-10137.

Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M. & Murzin, A.G. (1997) The solution structure of the s1 RNA binding domain: a member of an ancient nucleic acid-binding fold. Cell 88, 235-242.

Caton, A.J. & Koprowski, H. (1990) Influenza-virus hemagglutinin-specific antibodies isolated from a combinatorial expression library are closely related to the immune-response of the donor. Proc. Natl. Acad. Sci. USA 87, 6450-6454.

Chang, C.N., Landolfi, N.F. & Queen, C. (1991) Expression of antibody fab domains on bacteriophage surfaces - potential use for antibody selection. J. Immunol. 147, 3610-3614.

Clackson, T., Hoogenboom, H.R., Griffiths, A.D. & Winter, G. (1991) Making antibody fragments using phage display libraries. Nature 352, 624-628.

Davies, J. & Riechmann, L. (1995) An antibody VH domain with a lox-Cre site integrated into its coding region: bacterial recombination within a single polypeptide chain. FEBS Lett. 377, 92-96.

Davidson, A.R. & Sauer, R.T. (1994) Folded proteins occur frequently in libraries of random amino-acid-sequences. Proc. Natl. Acad. Sci. USA 91, 2146-2150.

Devereux, J., Haeberlie, P. & Smithies O. (1984) A comprehensive set of sequence analysis program for the VAX. Nucl. Acids Res. 12, 387-395.

Dower, W.J. & Fodor, S.P.A. (1991) The search for molecular diversity .2. recombinant and synthetic randomized peptide libraries. Annu Rep Med Chem 26, 271-280.

Edelhoch, H. & Osborne, J.C., Jr. (1976) The thermodynamic basis of the stability of proteins, nucleic acids, and membranes. Adv. Prot. Chem. 30, 183-250.

Eggertsson, G. & Söll, D. (1988) Transfer ribonucleic acid-mediated suppression of termination codons in *Escherichia Coli*. Microbiol. Rev. 52, 354-374.

Feng, D.F. & Dolittle. R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. of Molec. Evol. 25, 351-360.

Fincuane, M.D., Tuna, M., Lees, J.H. & Woolfson, D.N. (1999) Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. Biochemistry 38, 11604-11612.

Fire, A. & Xu, S.Q. (1995) Rolling replication of short dna circles. Proc. Natl. Acad. Sci. USA 92, 4641-4645.

Fodor. S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. & Solas. D. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251. 767-773.

Fontana, A., Laureto, P. de, Filipis, V. de, Scaramella, E. & Zambonin. M. (1997) Probing the partly folded states of proteins by limited proteolysis. Fold. Des. 2. R17-R26.

Gibson, T.J. (1984) Ph.D. Thesis, University of Cambridge, UK.

Goldstein, J., Pollitt, N.S. & Inouye, M. (1990) Major cold shock protein of Escherichia coli. Proc. Natl. Acad. Sci. USA 87, 283-287.

Greenfield, N & Fasman, G.D. (1969) Computed circular dichroism spectra for the evaluation of protein conformation. Biochemistry 8, 4108-4116.

Hardies, S.C., Hillen, W., Goodman, T.C. & Wells, R.D. (1979) High resolution thermal denaturation analyses of small sequenced DNA restriction fragments containing Escherichia coli lactose genetic control loci. J. Biol. Chem. 254, 5527-5534.

Hartley, R.W. (1993) Directed mutagenesis and barnase-barstar recognition. Biochemistry 32, 5978-5984.

Hawkins, R.E., Russell, S.J. & Winter, G. (1992) Selection of phage antibodies by binding-affinity - mimicking affinity maturation. J. Mol. Biol. 226, 889-896.

Hawkins, R.E. & Winter, G. (1992) Cell selection-strategies for making antibodies from variable gene libraries - trapping the memory pool. Eur. J. Immunol. 22, 867-870.

Hecht, M. (1994) De novo design of β-sheet proteins. Proc. Natl. Acad. Sci. USA 91, 8729-8730.

Higgins, D.G. & Sharp, P.M. (1989) Fast and sensitive multiple sequence alignment on a microcomputer. CABIOS 5, 151-153.

Hoogenboom, H.R., Griffiths, A.D., Johnson, K.S., Chiswell, D.J., Hudson, P. & Winter, G. (1991) Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. Nucleic Acids Res. 19, 4133-4137.

Hubbard, S.J., Eisenmenger, F. & Thornton, J.M. (1994) Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. Protein Science 3, 757-768.

Huse, W.D., Sastry, L., Iverson, S.A., Kang, A.S., Altingmees, M., Burton, D.R., Benkovic, S.J. & Lerner, R.A. (1989) Generation of a large combinatorial library of the immunoglobulin repertoire in phage-lambda. Science 246 1275-1281.

Hutchens, T. W., and Yip, T-T. (1993) New desorption strategies for the mass spectrometric analysis of macromolecules. Rapid Commun. Mass Spectrom. 7, 576-580.

Hutchison III, C.A., Phillips, S., Edgell, M.H., Gillam, S., Jahnke, P. & Smith, M. (1978) Mutagenesis at a specific position in a DNA sequence. J. Biol. Chem. 253, 6551-6560.

Jiang, W.N., Hou, Y. & Inouye, M. (1997) CspA, the major cold-shock protein of Escherichia coli, is an RNA chaperone. J. Biol. Chem. 272, 196-202.

Johnson, W.C. Jr (1990) Protein secondary structure and circular-dichroism - a practical guide. Proteins 7, 205-214.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M. (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262, 1680-1685.

Kang, A.S., Jones, T.M. & Burton, D.R. (1991) Antibody redesign by chain shuffling from random combinatorial immunoglobulin libraries. Proc. Natl. Acad. Sci. USA 88, 11120-11123.

Kristensen, P. & Winter, G. (1997) Proteolytic selection for protein folding using filamentous bacteriophages. Folding Des. 3, 321-328.

Kortemme, T., Ramirez-Alvardo, M. & Serrano, L. (1998) Design of a 20-amino acid, three-stamded β-sheet protein. Science 281, 253-256.

Lerner, R.A., Kang, A.S., Bain, J.D., Burton, D.R. & Barbas, C.F. (1992) Antibodies without immunization Science 258, 1313-1314.

Low, N.M., Holliger, P. & Winter, G. (1996) Mimicking somatic hypermutation: Affinity maturation of antibodies displayed on bacteriophage using a bacterial. J. Mol. Biol. 260, 359-368.

Lowman HB, Bass SH, Simpson N. & Wells, J.A. (1991) Selecting high-affinity binding-proteins by monovalent phage display. Biochemistry 30, 10832-10838.

Lubienski, M.J., Bycroft, M., Jones, D.N.M. & Fersht, A.R. (1993) Assignment of the backbone H-1 and N-15 nmr resonances and secondary structure characterisation of barstar. FEBS Lett. 332, 81-87.

Marks, J.D., Hoogenboom, H.R., Bonnert, T.P., McCafferty, J., Griffiths, A.D. & Winter, G. (1991) By-passing immunization - human-antibodies from v-gene libraries displayed on phage. J. Mol. Biol. 222, 581-597.

Marks, J.D., Hoogenboom HR, Griffiths AD, Winter G (1992) Molecular evolution of proteins on filamentous phage - mimicking the strategy of the immune-system. J. Biol. Chem. 267, 16007-16010.

McCafferty, J., Griffiths, A.D., Winter, G. & Chiswell, D.J. (1990) Phage antibodies - filamentous phage displaying antibody variable domains. Nature 348, 552-554.

Meiering, E.M., Serrano, L. & Fersht, A.R. (1992) Effect of active site residues in barnase on activity and stability. J. Mol. Biol. 225, 585-589.

Mullinax, R.L., Gross, E.A., Amberg, J.R., Hay, B.N., Hogrefe, H.H., Kubitz, M.M., Greener, A., Altingmees, M., Ardourel, D., Short, J.M., Sorge, J.A. & Shopes, B. (1990) Identification of human-antibody fragment clones specific for tetanus toxoid in a bacteriophage-lambda immunoexpression library. Proc. Natl. Acad. Sci. USA 87, 8095-8099.

Myers, E.W. & Miller, W. (1988) "Optimal Alignments in Linear Space", CABIOS 4, 11-17.

Needleman, S.B. & Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48, 444-453.

Newkirk, K., Feng, W.Q., Jiang, W.N., *et al.*(1994) Solution nmr structure of the major cold shock protein (cspa) from escherichia-coli - identification of a binding epitope for DNA. Proc. Natl. Acad. Sci. USA 91, 5114-5118.

Pace, C.N. (1990) Conformational stability of globular proteins. Trends Biochem. Sci. 15, 14-17.

Persson, M.A.A., Caothien, R.H. & Burton, D.R. (1991) Generation of diverse high-affinity human monoclonal-antibodies by repertoire cloning. Proc. Natl. Acad. Sci. USA 88, 2432-2436.

Piotto, M. Saudek, V. & Sklenar, V. (1992) Gradient-tailored excitation for single-quantum nmr-spectroscopy of aqueous-solutions. J. Biomolecular NMR 2, 661-665.

Quinn, T.P., Tweedy, N.B., Williams, R.W., Richardson, J.S. & Richardson, D.C. (1994) Betadoublet: De novo design, synthesis, and characterisation of a β-sandwich protein. Proc. Natl. Acad. Sci. USA 91, 8747-8751.

Rakonjac, J., Jovanovic, G. & Model, P. (1997) Filamentous phage infection-mediated gene expression: construction and propagation of the gIII deletion mutant helper phage R408d3. Gene 198, 99-103.

Regan, L. (1998) Proteins to order? Structure 6, 1-4.

Riechmann, L. & Davies, J. (1995) Backbone assignment, secondary structure and Protein A binding of an isolated, human antibody VH domain. J. Biomol. NMR 6, 141-152.

Riechmann, L. & Holliger, P. (1997) The C-terminal doamin of TolA is the coreceptor for filamentous phage infection of E. coli. Cell 90, 351-360.

Riechmann, L., & Weill, M. (1993) Phage display and selection of a site-directed randomized single-chain antibody Fv fragment for its affinity improvement. Biochemistry 32, 8848-8855.

Sauer, R.T. (1996) Protein folding from a combinatorial perspective. Folding Des. 1, R27-R30.

Schindelin, H., Maraheil, M.A. & Heinemann, U. (1994) Crystal structure of CspA, the major cold shock protein of Escherichia coli. Proc. Natl. Acad. Sci. USA 91, 5119-5123.

Schroder, K., Graumann, P., Schnuchel, A., Holak, T.A. & Marahiel, M.A. (1995) Mutational analysis of the putative nucleic acid-binding surface of the cold-shock domain, cspb, revealed an essential role of aromatic and basic residues in binding of single-stranded-dna containing the y-box motif. Mol. Microbiol. 16, 699-708.

Scott, J.K. & Smith. G.P. (1990) Searching for peptide ligands with an epitope library. Science 249. 386-390.

Sieber, V., Plueckthun, A. & Schmid, F.X. (1998) Selecting proteins with improved stability by a phage-based method. Nat. Biotechnol. 16. 955-960.

Smith, T.F. & Waterman, M.S. (1981) Comparison of Bio-sequences. Advances in Applied Mathematics 2, 482-489.

Smith, T.F., Waterman, M.S. & Sadler, J.R. (1983) Statistical characterisation of nucleic acid sequence functional domains. Nucleic Acids Res. 11, 2205-2220.

Sternberg, N. & Hamilton, D. (1981) Bacteriophage P1 site-specific recombination. I. Recombination between loxP sites. J. Mol. Biol., 150, 467-486.

Tame J.R., Murshudov, G.N., Dodson, E.J., Neil, T.K., Dodson, G.G., Higgins, C.F. & Wilkinson, A.J. (1994) The structural basis of sequence-independent peptide binding by OppA protein. Science 264, 1578-1581.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) ClusterW: improving the sensitivity of progressive multiple sequence alignment through sequence weighing, positions-specific gap penalties and weight matrix choice. Nucleic Acid Res. 22, 4673.

Wilbur, W.J. & Lipman, D.J. (1983) Rapid similarity searches of nucleic-acid and protein data banks. Proc. Natl. Acad. Sci. USA 80, 726-730.

Wüthrich, K. (1986) NMR of proteins and nucleic acids. Chapter 3. Wiley & Sons.

Zacher, A.N., Stock, C.A., Golden, J.W. & Smith, G.P. (1980) A new filamentous phage cloning vector: fd-tet. Gene 9, 127-140.

Cited patents and patent applications:

PCT/GB00/00030
PCT/GB98/01889
WO84/03564
WO88/08453
WO91/05058
WO90/05785
WO90/07003
WO90/15070
WO91/02076
WO92/00091
WO92/02536
WO92/10092
WO93/06121
WO95/11922
WO95/22625
U.S. Patent No. 4,631,211
U.S. Patent No. 5,143,854

Claims

1. A chimaeric folded protein domain, when derived from a repertoire of chimaeric proteins, comprising two or more sequence segments derived from parent amino acid sequences that are non-homologous.

2. A chimaeric folded protein according to claim 1, wherein two or more of the sequence segments are combined non-covalently.

3. A chimaeric folded protein domain according to claim 1, in which at least one of the parent amino acid sequences is derived from a protein.

4. A chimaeric folded protein domain according to claim 3, in which at least one of the parent amino acid sequences is derived from a protein selected from the group consisting of a naturally occurring protein, an engineered protein, a protein with a known binding activity, a protein with a known binding activity for an organic compound, a protein with a known binding activity for a peptide or polypeptide, a protein with a known binding activity for a carbohydrate, a protein with a known binding activity for a nucleic acid, a known binding activity for a hapten, a protein with a known binding activity for a steroid, a protein with a known binding activity for an inorganic compound, and a protein with an enzymatic activity.

5. A chimaeric folded protein domain according to claim 1, in which the parent amino acid sequences are derived from the open reading frames of a single genome or part thereof:

(a) wherein said reading frames are the natural reading frame of the genes; or

(b) wherein said reading frames are not the natural reading frame of the genes.

6. A chimaeric folded protein domain according to claim 1, in which the parent amino acid sequences are derived from the open reading frames of two or more genomes or part thereof:

(a) wherein said reading frames are the natural reading frame of the genes; or

(b) wherein said reading frames are not the natural reading frame of the genes.

7. A chimaeric protein domain according to claim 1, which is resistant to *in vivo* or *in vitro* proteolysis by protease enzymes.

8. A chimaeric protein according to claim 1, wherein the sequence segments originate from parent domains with the same polypeptide fold in at least part of the structure.

9. A chimaeric protein according to claim 1, wherein the sequence segments originate from parent domains with different polypeptide folds in at least part of the structure.

10. A chimaeric protein domain according to claim 1, which has a free energy of folding greater than 1.6 kcal/mol.

11. A chimaeric protein domain according to claim 10, which has a free energy of folding greater than 3 kcal/mol.

12. A chimaeric protein domain according to claim 11, which has a free energy of folding of greater than 5 kcal/mol.

13. A chimaeric folded protein according to any preceding claim, wherein one or more of the sequence segments is fused to one or more additional and complete protein domains.

14. A chimaeric protein domain according to claim 1 fused to the coat protein of a filamentous bacteriophage, said bacteriophage encapsidating a nucleic acid encoding the protein domain.

15. A chimaeric protein domain according to claim 1, wherein a single sequence segment originates from human proteins.

16. A chimaeric protein domain according to claim 1. wherein two or more sequence segments originate from human proteins.

17. A chimaeric protein domain according to claim 15 or claim 16. wherein at least one of the segments is derived from a source other than a human protein.

18. A chimaeric protein domain according to claim 17, wherein all segments are derived from human proteins.

19. A chimaeric protein according to claim 1 comprising an epitope of the parent amino acid sequence.

20. A chimaeric protein according to claim 19 comprising a conformational epitope.

21. A chimaeric protein according to claim 1 that cross-reacts with antibodies raised against a parent amino acid sequence.

22. A chimaeric protein according to claim 1 that cross-reacts with antibodies raised against the folded parent protein.

23. A chimaeric protein according to claim 1, for use in vaccination against one or more of the amino acid sequences from which the chimaera is derived.

24. A chimaeric protein according to claim 1, for administration to a human for therapeutic purposes.

25. A chimaeric protein according to claim 1, for use in a commercial product to which humans are exposed.

26. A chimaeric protein according to claim 1, wherein the amino acid sequences are altered to increase stability or function of the chimaeric protein.

27. A chimaeric nucleic acid encoding a protein domain according to claim 1.

28. A method for preparing a protein domain according to claim 1, comprising the steps of:

(a) providing a first library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences:

(b) providing a second library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences:

(c) combining the coding sequences to form a combinatorial library of nucleic acids. said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first and second libraries;

(d) transcribing and/or translating the contiguous coding sequences to produce the encoded protein domains;

(e) selecting the chimaeric protein domains which are able to adopt a folded structure or to fulfil a specific function;

29. A method according to claim 28, further comprising the steps of:

(f) analysing the sequence of the selected chimaeric protein domains to identify the origins of the sequence segments; and

(g) comparing the sequences of each of the parent amino acid sequences to identify whether the sequences of the parent amino acid sequences are non-homologous.

30. A method for preparing a protein domain according to claim 9 or claim 10, comprising the steps according to claim 28 or claim 29 and the additional step of:

(h) comparing the structures of each of the parent amino acid sequences to identify whether they have the same polypeptide folds in whole or in part.

31. A method according to claim 28 , wherein step (b) and (c) are modified as follows:

(b) providing a partner coding sequence encoding a sequence segment derived from one protein;

(c) combining the library and partner coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first library and the partner coding sequence.

32. A method according to claim 28, wherein the domains which are able to adopt a folded structure are selected by one or more methods selected from the group consisting of *in vivo* proteolysis, *in vitro* proteolysis, binding ability, functional activity and expression.

33. A method according to claim 32, wherein said binding ability is to an antibody raised against a parent protein

34. A method for preparing a protein domain according to claim 26, wherein the sequence segments of the parent amino acid sequences are altered subsequent to their juxtaposition, comprising one or more of the following steps:

(a) designing and introducing specific or random mutations at predefined positions within the gene of the chimaeric protein;

(b) deletion of nucleotides within the gene of the chimaeric protein so as to delete amino acid residues;

(c) insertion of nucleotides within the gene of the chimaeric protein so as to insert amino acid residues

(d) appending nucleotides to the gene of the chimaeric protein so as to append amino acid residues;

(e) randomly introducing mutations in all or part of the gene encoding the chimaeric protein through recombinant DNA technology;

(f) randomly introducing mutations in the gene of the chimaeric protein through propagation in mutating cells;

(g) introduction of derivatives of natural amino acid during chemical synthesis;

(h) chemical derivatisation of amino acid groups after synthesis;

(i) multimerisation of the chimaeric proteins through concatenation of two or more copies of the gene in a single open reading frame;

(j) multimerisation of the chimaeric proteins through covalent linkage of two or more copies of the chimaeric protein domain after translation;

(k) multimerisation of the chimaeric proteins through fusion to a multimeric partner.

35. A chimaeric protein domain according to claim 1, comprising at least one reaction group for covalent linkage.

36. A chimaeric protein domain according to claim 1, comprising at least one reaction group for non-covalent linkage.

37. A chimaeric protein domain according to claim 1, comprising at least one D-amino acid.

38. A chimaeric protein domain according to claim 1, comprising at least one non-naturally-occurring amino acid.

39. A chimaeric protein domain according to claim 1, comprising at least one amino acid having a label or a tag.

40. A chimaeric folded protein domain when derived from a repertoire of chimaeric folded proteins, comprising two or more sequence segments derived from parent amino acid sequences wherein each of said segments comprises common sequences in the chimaeric protein, and in which said common sequences are not designed or selected to consist solely of one or more complete structural elements.

41. A folded chimaeric protein domain according to claim 40, in which the region of common sequence is 10 or more identical amino acid residues in length.

42. A folded chimaeric protein domain according to claim 41, in which the region of common sequence is 20 or more identical amino acid residues in length.

43. A chimaeric folded protein domain when derived from a repertoire of chimaeric folded proteins, comprising two or more sequence segments wherein each of said segments:

(a) is derived from parent proteins with a common fold; and

(b) comprises a common region of the common fold and in which said common region of the fold is not designed or selected to consist of one or more complete structural elements.

44. A chimaeric folded protein domain according to claim 43 wherein each of said segments is derived from different proteins which are homologous in sequence.

45. A chimaeric folded protein domain according to claim 43 wherein each of said segments is derived from the same protein.

46. A folded chimaeric protein domain according to claim 43 in which the common region of the fold is 10 or more amino acid residues in length.

47. A folded chimaeric protein domain according to claim 43 which the common region of the fold is 20 or more amino acid residues in length.

48. A chimaeric folded protein domain according to claim 43, in which the amino acid sequences of the parent proteins are derived from the open reading frames of a genome or part thereof, wherein said reading frames are the natural reading frame of the genes.

49. A chimaeric protein domain according to claim 40, which is resistant to *in vivo* or *in vitro* proteolysis by protease enzymes.

50. A chimaeric protein domain according to claim 40, which has a free energy of folding greater than 1.6 kcal/mol.

51. A chimaeric folded protein according to claim 40, wherein one or more of the sequence segments is fused to one or more additional and complete protein domains.

52. A chimaeric protein domain according to claim 40 fused to the coat protein of filamentous bacteriophage, said bacteriophage encapsidating a nucleic acid encoding the protein domain.

53. A chimaeric protein domain according to claim 40, wherein a single sequence segment originates from a human protein.

54. A chimaeric protein domain according to claim 40, wherein two or more of the sequence segments originate from a human protein.

55. A chimaeric protein domain according to claim 40, wherein at least one of the segments is derived from a source other than a human protein.

56. A chimaeric protein domain according to claim 47, wherein all segments are derived from human proteins.

57. A chimaeric protein according to claim 40 comprising an epitope of the parent amino acid sequence.

58. A chimaeric protein according to claim 57 comprising a conformational epitope.

59. A chimaeric protein according to claim 40 that cross-reacts with antibodies raised against a parent amino acid sequence.

60. A chimaeric protein according to claim 40 that cross-reacts with antibodies raised against the folded parent protein.

61. A chimaeric protein according to claim 40, for use in vaccination against the parent protein(s) from which the chimaera is derived.

62. A chimaeric protein according claim 40, for administration to a human for therapeutic purposes.

63. A chimaeric protein according to claim 40, for use in a commercial product to which humans are exposed.

64. A chimaeric protein according to claim 40, wherein the amino acid sequences are altered to increase stability or function of the chimaeric protein.

65. A chimaeric nucleic acid encoding a protein domain according to claim 40.

66. A method for preparing a chimaeric protein domain according to claim 40, comprising the steps of:

(a) providing a first library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences;

(b) providing a second library of nucleic acids, said library comprising coding sequences encoding sequence segments derived from one or more amino acid sequences;

(c) combining the coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first and second libraries;

(d) transcribing and/or translating the contiguous coding sequences to produce the encoded protein domains; and

(e) selecting the chimaeric protein domains, which are able to adopt a folded structure or to fulfil a specific function.

67. A method according to claim 66, further comprising the steps of:

(f) analysing the sequence of the selected chimaeric protein domains to identify the origins of the sequence segments; and

(g) comparing the sequences to identify whether they comprise common sequences according to claim 40.

68. A method according to claim 66 for preparing a chimaeric protein domain according to claim 44, wherein step (g) is replaced and step (h) is added such that:

(g) comparing the structures of the parent amino acid sequences to identify whether the parent amino acid sequences have a common fold; and

(h) identifying whether the segments comprise a common region of the common fold.

69. A method according to claim 66, wherein step (b) and (c) are modified such that:

(b) providing a partner coding sequence encoding a sequence segment derived from one protein;

(c) combining the library and partner coding sequences to form a combinatorial library of nucleic acids, said nucleic acids comprising contiguous coding sequences encoding sequence fragments derived from the first library and the partner coding sequence.

70. A method according to claim 66 wherein the domains which are able to adopt a folded structure are selected by one or more methods selected from the group consisting of *in vivo* proteolysis, *in vitro* proteolysis, binding ability, functional activity and expression.

71. A method according to claim 70, wherein said binding ability is to an antibody raised against a parent protein

72. A method for preparing a chimaeric protein domain according to claim 40, wherein the sequence segments of the parent amino acid sequences are altered subsequent to their juxtaposition, comprising one or more of the following steps:

(a) designing and introducing specific or random mutations at predefined positions within the gene of the chimaeric protein;

(b) deletion of nucleotides within the gene of the chimaeric protein so as to delete amino acid residues;

(c) insertion of nucleotides within the gene of the chimaeric protein so as to insert amino acid residues

(d) appending nucleotides to the gene of the chimaeric protein so as to append amino acid residues;

(e) randomly introducing mutations in all or part of the gene encoding the chimaeric protein through recombinant DNA technology;

(f) randomly introducing mutations in the gene of the chimaeric protein through propagation in mutating cells;

(g) introduction of derivatives of natural amino acid during chemical synthesis;

(h) chemical derivatisation of amino acid groups after synthesis;

(i) multimerisation of the chimaeric proteins through concatenation of two or more copies of the gene in a single open reading frame;

(j) multimerisation of the chimaeric proteins through covalent linkage of two or more copies of the chimaeric protein domain after translation;

(k) multimerisation of the chimaeric proteins through fusion to a multimeric partner.

73. A chimaeric protein domain according to claim 40, comprising at least one reaction group for covalent linkage.

74. A chimaeric protein domain according to claim 40, comprising at least one reaction group for non-covalent linkage.

75. A chimaeric protein domain according to claim 40, comprising at least one D-amino acid.

76. A chimaeric protein domain according to claim 40, comprising at least one non-naturally-occurring amino acid.

77. A chimaeric protein domain according to claim 40, comprising at least one amino acid having a label or a tag.

Figure 1

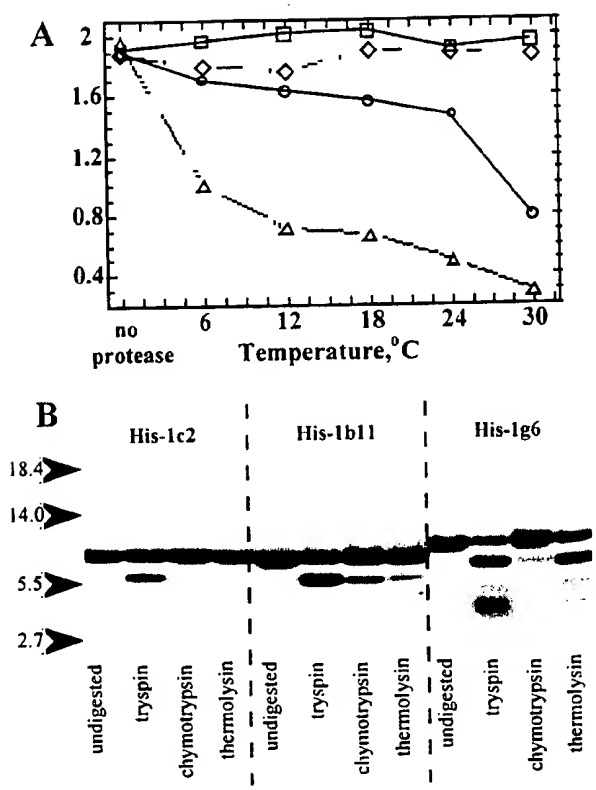
A
2
1.6
1.2
0.8
0.4
no protease
6
12
18
24
30
Temperature,°C
B
His-1c2
His-1b11
His-1g6
18.4
14.0
5.5
2.7
undigested
tryspin
chymotrypsin
thermolysin
undigested
tryspin
chymotrypsin
thermolysin
undigested
tryspin
chymotrypsin
thermolysin

Figure 2

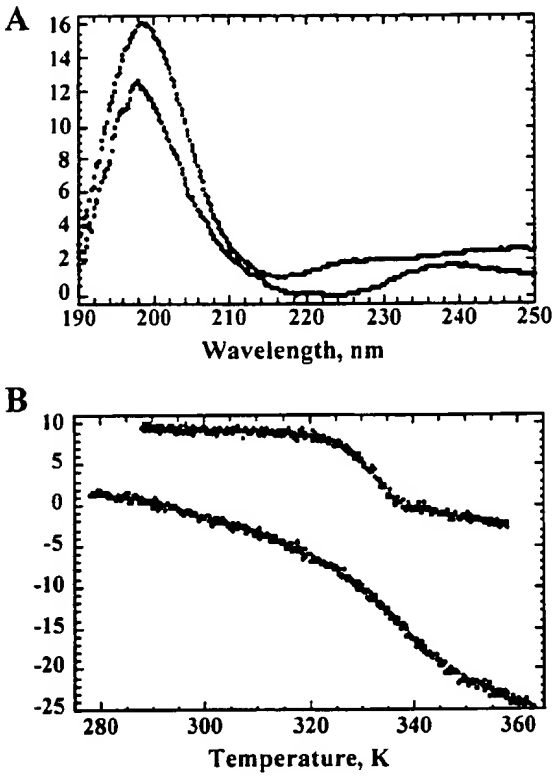
A
16
14
12
10
8
6
4
2
0
190
200
210
220
230
240
250
Wavelength, nm
B
10
5
0
-5
-10
-15
-20
-25
280
300
320
340
360
Temperature, K

Figure 3

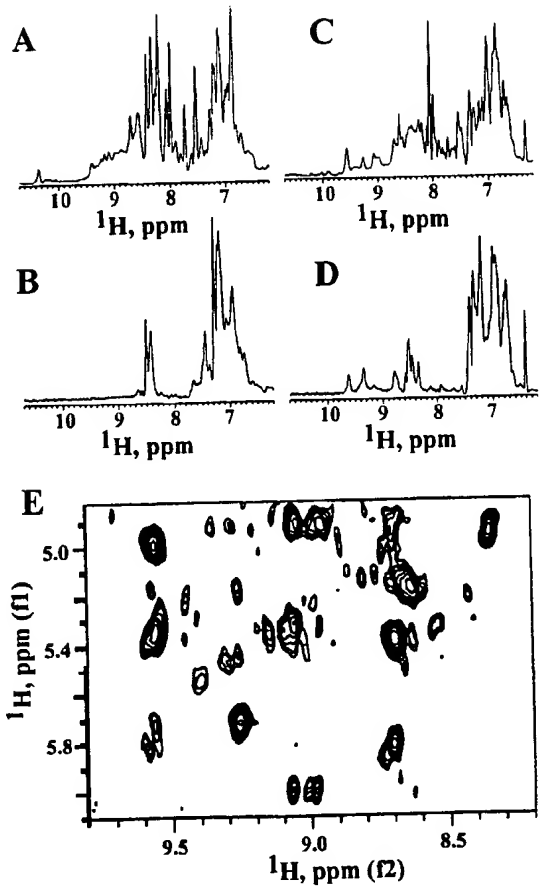

Figure 4